BASIL-DCM: Biophysical Amortized Scalable Inference for Latent Dynamic Causal Modeling

Moein Khajehnejad[1,3*] **Forough Habibollahi**[2,*] **Leonardo Novelli**[1,3] **Adeel Razi**[1,3,4,5]
[1]Turner Institute for Brain and Mental Health, School of Psychological Sciences, Monash University, Australia
[2]Tether Evo [3]Monash Biomedical Imaging, Monash University, Australia
[4]Queen Square Institute of Neurology, UCL, UK
[5]CIFAR Global Scholars Program, Toronto, Ontario, Canada [*]Equal contribution

Abstract

Estimating directed, weighted, and signed interactions among brain regions (*effective connectivity*) from fMRI requires disentangling neural dynamics from delayed and nonlinear hemodynamic transformations. Dynamic Causal Modeling (DCM) provides a principled solution by inverting a biophysical generative model, but the standard Variational Laplace inversion method is computationally prohibitive for large parcellations and cohort-scale datasets. We introduce *Biophysical Amortized Scalable Inference for Latent Dynamic Causal Modeling* (**BASIL-DCM**), a physics-informed amortized inference model that estimates subject-specific effective connectivity and biophysical DCM parameters, including ROI-wise hemodynamic transit time, spectral properties of endogenous neural fluctuations, and observation noise, *in a single forward pass*. BASIL-DCM combines a linear-time state-space temporal encoder with an ROI-wise Transformer to capture long-range temporal dependencies and inter-regional interactions. The model is trained on Human Connectome Project-informed simulations generated by numerically integrating the coupled neuronal and nonlinear hemodynamic DCM equations, with effective connectivity initialized from regression DCM and complementary biophysical parameters sampled from physiologically plausible ranges. Learning is additionally constrained by a differentiable spectral-consistency objective based on the resting-state linearization used in spectral DCM. This approach enables fast, uncertainty-aware cortex-wide network inference while preserving mechanistic interpretability and ensuring consistency with biophysical dynamics.

1 Introduction

Modelling causal influence among brain regions is central to mechanistic neuroscience. *Effective connectivity* formalizes this as a network of directed, weighted, and signed connections (representing excitatory/inhibitory influences), where each connectivity parameter quantifies the rate of change in neuronal population activity in response to another. By modeling interactions between latent (unobserved) neuronal populations, this approach moves beyond popular measures of statistical dependence among observed neuroimaging signals, e.g. functional connectivity [1, 2]. However, estimating effective connectivity from functional magnetic resonance imaging (fMRI) is challenging because the observed blood-oxygen-level-dependent signals (BOLD) are delayed, and involve nonlinear hemodynamic transformations of neural activity [3, 4, 5]. Dynamic Causal Modeling (DCM) addresses this by specifying a hierarchical biophysical generative model of neuronal dynamics and hemodynamics, then performing Bayesian model inversion to infer the latent connectivity and biophysical parameters [6, 5]. Despite its interpretability, classical DCM is difficult to scale: inversion is performed separately for each subject, and memory usage and computational complexity grow polynomially as network size grows. This conflicts with modern neuroimaging practice, where parcellations often contain tens to hundreds of regions [7] and cohort studies include thousands of subjects [8, 9, 10].

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

We address this bottleneck with *amortized inference*. Rather than solving a (costly) optimization problem for each subject, we train a neural inverse model so that inference requires only a single forward pass. This idea is related to simulation-based inference and neural posterior estimation, which learn conditional distributions over parameters without requiring a tractable likelihood [11, 12, 13]. Training is performed using simulations informed by empirical resting-state fMRI data from the Human Connectome Project (HCP) [9]. Specifically, we use a scalable version of DCM (regression DCM [14]; rDCM) to obtain realistic estimates of effective connectivity, which serve as part of the ground-truth. Because rDCM does not infer certain biophysical parameters of interest (e.g., hemodynamics and endogenous neural fluctuation parameters), we complement these with values randomly sampled from physiologically realistic ranges. The full parameter set is then used to numerically integrate the neuronal dynamics and nonlinear Balloon–Windkessel equations to generate the simulated BOLD time series. This time-domain simulation is distinct from the analytic spectral operator used for regularization, which linearizes the hemodynamic equations around the resting state following spectral DCM. This procedure ensures that the training data cover a realistic region of parameter space and is intended to reduce the simulation-to-real gap.

Two considerations shape the model design. First, resting-state fMRI timeseries are usually long ($T \sim 10^3$), making time-wise self-attention costly due to $O(T^2)$ complexity. Second, purely black-box approaches may fit data features while violating dynamical and spectral properties required for biological plausibility. Spectral DCM highlights that the frequency spectrum of resting-state activity, specifically the cross-spectral density (CSD), is strongly constrained by endogenous fluctuations with a power-law spectrum (e.g. $1/f$) [15, 16, 17].

We introduce *Biophysical Amortized Scalable Inference for Latent Dynamic Causal Modeling* , **BASIL-DCM** (hereafter, BASIL), a physics-informed amortized inference framework for scalable cortex-wide effective connectivity estimation from fMRI. BASIL combines a linear-time selective state-space temporal encoder with an ROI-wise Transformer across regions. The temporal encoder summarizes long BOLD sequences efficiently, while the spatial module estimates a directed connectivity matrix and associated DCM parameters, including global spectral parameters and ROI-wise hemodynamic and noise terms such as transit time and region-specific observation noise—capabilities largely absent from modern scalable baselines that estimate only connectivity [18]. These parameters are essential for mechanistic interpretation, as they disentangle neuronal interactions from hemodynamic and noise processes that vary across regions and subjects [6, 5]. Moreover, spectral and hemodynamic features have clinical value as they change with aging, drug use, neurodegenerative and psychiatric diseases [19, 20]. We train BASIL on HCP-informed simulated resting-state BOLD time series and regularize learning with a differentiable spectral-consistency objective derived from spectral DCM [15, 21, 17]. BASIL enables fast cortex-wide network inference with learned edge-wise variance estimates while preserving the mechanistic interpretability that motivates DCM. Once trained, BASIL returns point estimates of the directed connectivity and associated biophysical parameters, together with a variance estimate for each directed connectivity edge, in a single forward pass. Moreover, the pretrained model can be, in principle, reused across datasets, requiring only lightweight adaptation via the spectral consistency objective to account for dataset-specific BOLD dynamics and acquisition characteristics.

2 Related Work

DCM and spectral DCM. DCM frames effective connectivity estimation as Bayesian inversion of a directed neuronal state-space model with a biophysical hemodynamic observation model [6, 5]. Spectral DCM adapts this framework to resting-state by fitting cross-spectral density (CSD) statistics under stationarity assumptions and modeling scale-free endogenous fluctuations [15, 16, 17]. These approaches are mechanistically interpretable, but per-subject iterative inversion limits scalability.

Amortized and likelihood-free inference. Likelihood-free methods such as Simulation-based Inference and neural posterior estimation learn conditional distributions over parameters from paired observations, enabling fast amortized inference without explicit likelihood evaluation [11, 12, 13]. BASIL follows this amortized-inference perspective, but is trained on HCP-informed data using DCM-derived targets and constrained by analytic spectral regularization.

Scalable effective connectivity. Several approaches have sought to make effective-connectivity estimation more scalable. rDCM improves scalability by forgoing the strict separation between latent and observed variables and the regional hemodynamic variation and assuming no covariance among the latent parameters [14]. Recent data-driven approaches have also used surrogate dynamical models and perturbation-based identification. NPI, for example, computes causal influence by virtually

perturbing a learned surrogate brain [22]. While scalable, NPI does not invert a DCM biophysical observation model or estimate DCM hemodynamic and spectral-noise parameters. BASIL instead targets a DCM-style parameterization, inferring both directed couplings and biophysical parameters while enforcing spectral plausibility through an analytic CSD-based regularizer.

Causal discovery from fMRI time series. Directed influences are often estimated using autoregressive criteria, constraint-based methods, and conditional-independence testing [23, 24, 25]. Whole-brain causal discovery in fMRI remains difficult due to hemodynamic filtering and confounds, motivating recent methods for latent lag structures [26]. BASIL differs by embedding directed inference in a biophysical generative model and learning an amortized inverse that outputs both connectivity and biophysical parameters.

Long-context sequence models for brain signals. Transformers provide flexible sequence modeling but incur $O(T^2)$ time-wise attention cost. State-space sequence models offer an efficient alternative with linear-time recurrence and strong long-context performance [27]. In parallel, large-scale representation learning for brain signals continues to develop, including foundation-model-style efforts [28, 29]. BASIL leverages state-space modeling for efficiency, but focuses on mechanistic inversion under explicit spectral constraints rather than purely predictive representation learning.

3 Method

We propose BASIL , a physics-informed amortized inverse model for estimating DCM parameters from resting-state fMRI. These parameters are $\theta = \{A, a, b, c, \tau\}$, where $A \in \mathbb{R}^{R\times R}$ is directed effective connectivity (i.e., an adjacency matrix or a Jacobian), $a = (a_1, a_2)$ parameterizes the amplitude and exponent of endogenous neuronal fluctuations, and $b = (b_1, b_2)$ and c parameterize global and ROI-wise observation noise. The hemodynamic response is described by an ROI-specific transit time τ, which captures regional BOLD delay differences. These parameters enter the DCM forward model through $\dot{x}(t) = Ax(t) + u(t)$, $G_u(\omega) = a_1\omega^{-a_2}$, $y_i(t) = \mathcal{H}_{\tau_i}[x_i](t) + e_i(t)$, and $G_{e_i}(\omega) = (b_1 + c_i)\omega^{-b_2}$, where $\mathcal{H}_{\tau_i}$ denotes the ROI-specific hemodynamic mapping. Full equations are provided in Appendix A. All parameters are subject-specific and summarized in Appendix A and Table S1. Estimating these parameters jointly helps distinguish variation in directed neuronal coupling from variation in endogenous neuronal fluctuations, observation noise, and regional hemodynamic delays, reducing the risk that non-neuronal BOLD variability is misattributed to effective connectivity. Classical DCM inversion solves a nonlinear subject-specific optimization problem, making large-scale inference computationally expensive. In contrast, BASIL learns a shared inverse operator

$$f_\phi : Y \mapsto \theta,$$

that maps observed BOLD time series (Y) directly to subject-specific DCM parameters (θ). Crucially, this inverse operator is not treated as a generic black-box regressor: its architecture mirrors the structure of the DCM forward model, separating local temporal dynamics, inter-regional 'directed' connectivity, subject-level variability, and spectral consistency. Given a batch of BOLD signals:

$$Y \in \mathbb{R}^{B\times R\times T},$$

where B, R, and T denote batch size, number of regions, and number of time points, BASIL jointly produces estimates for all spectral DCM parameters: $\hat{\theta} = \{\hat{A}, \hat{a}, \hat{b}, \hat{c}, \hat{\tau}\}$. Fig. 1 summarizes the framework. For full architecture and implementation details, see Appendix B and Tables S2–S3.

3.1 Structured Amortized Inversion

The DCM forward model couples temporal evolution within each region with directed interactions between regions. We therefore structure the inverse model around two complementary representations: a temporal representation of local ROI dynamics and a spatially contextualized representation of inter-regional connectivity. Specifically, BASIL computes

$$Z_{\text{temp}} = f_{\text{temp}}(Y), \quad Z_{\text{phase}} \quad = f_{\text{phase}}(\mathcal{H}(Y)), \quad Z_{\text{spat}} = f_{\text{spatial}}(Z_{\text{temp}}),$$

where $Z_{\text{temp}} \in \mathbb{R}^{B\times R\times d_{\text{time}}}$ denotes the temporal ROI embeddings, $Z_{\text{phase}} \in \mathbb{R}^{B\times R\times d_{\text{phase}}}$ denotes the phase-sensitive embeddings, and $Z_{\text{spat}} \in \mathbb{R}^{B\times R\times d_{\text{node}}}$ denotes the spatially contextualized ROI embeddings. Here, f_{temp} is a Mamba-based temporal encoder [27], $\mathcal{H}(Y)$ denotes the Hilbert-phase representation of the BOLD signals, f_{phase} is a multiscale phase encoder, and f_{spatial} is a Transformer over ROI tokens. The resulting representations are then routed to parameter-specific heads:

$$\hat{\theta} = f_{\text{head}}\left(Z_{\text{spat}}, Z_{\text{temp}}, Z_{\text{phase}}\right).$$

This factorization provides inductive bias by modeling how each region evolves over time, how regions interact, and constraining the resulting parameters through DCM generative physics.

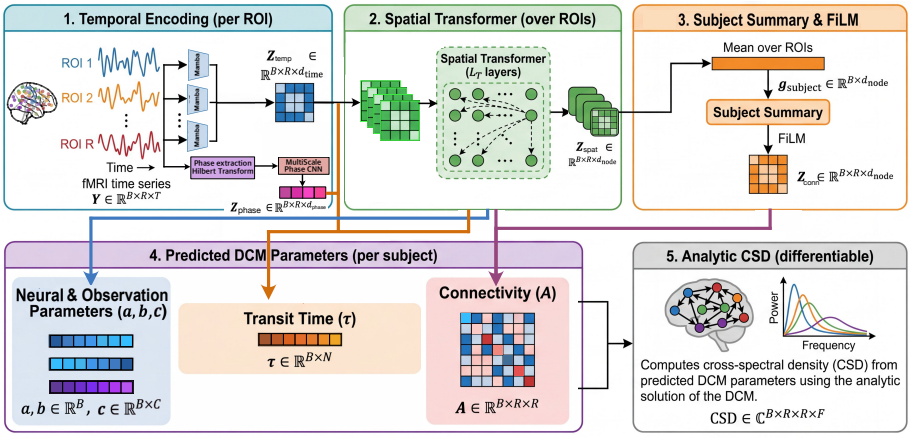


Figure 1: **Overview of BASIL .** BASIL amortizes DCM inversion by mapping resting-state fMRI to subject-specific effective connectivity and biophysical parameters. ROI time series are first encoded by a Mamba-based temporal module with phase-aware timing features. A spatial Transformer then models adaptive inter-regional interactions. A subject-level FiLM module conditions connectivity prediction on global brain-state context. Estimated DCM parameters are passed through a differentiable CSD module, enforcing consistency with the DCM forward model in the spectral domain.

Temporal encoding of local dynamics. The temporal encoder maps the BOLD trajectories into ROI-wise latent representations, $Z_{\text{temp}} \in \mathbb{R}^{B\times R\times d_{\text{time}}}$ (Fig. 1.1). We instantiate f_{temp} via a selective state-space model (Mamba). This choice is motivated by the inverse nature of the problem: effective connectivity is observed only indirectly through delayed and slow hemodynamic responses, requiring long-context temporal representations that remain scalable across subjects and ROIs. In parallel, we extract phase-sensitive features, $Z_{\text{phase}} \in \mathbb{R}^{B\times R\times d_{\text{phase}}}$, from the Hilbert-phase representation of BOLD signals using f_{phase} as a multiscale temporal CNN. These features provide explicit sensitivity to oscillatory timing structure and are used primarily for estimating hemodynamic transit times.

Spatial modeling of directed coupling. DCM's main feature is the inference of directed interactions between regions. To model these interactions, BASIL first projects the temporal ROI embeddings from d_{time} to d_{node}, and then applies a spatial Transformer, f_{spatial}, over the projected ROI tokens (Fig. 1.2), yielding $Z_{\text{spat}} \in \mathbb{R}^{B\times R\times d_{\text{node}}}$. Self-attention [30] allows each ROI to condition its representation on every other ROI, providing a flexible interaction operator. Because effective connectivity is signed, asymmetric, and subject-dependent, this Transformer mechanism provides the necessary adaptive capacity to learn latent connectivity structure from BOLD dynamics.

Subject-conditioned connectivity representations. Effective connectivity varies across individuals through both local ROI dynamics and global brain-state context. We therefore summarize each subject by applying a subject-summary network f_{subj} to the mean-pooled spatial ROI embeddings (Fig. 1.3):

$$g_{\text{subj}} = f_{\text{subj}}\left(\frac{1}{R}\sum_{i=1}^{R} Z_{\text{spat},i}\right).$$

The subject embedding is then used to condition the spatial ROI embeddings through feature-wise linear modulation (FiLM) [31]:

$$Z_{\text{conn}} = f_{\text{FiLM}}(Z_{\text{spat}}, g_{\text{subj}}) = \gamma(g_{\text{subj}}) \odot Z_{\text{spat}} + \beta(g_{\text{subj}}),$$

where $\gamma(\cdot)$ and $\beta(\cdot)$ are learned feature-wise scale and shift functions (see Appendix B). We refer to $Z_{\text{conn}} \in \mathbb{R}^{B\times R\times d_{\text{node}}}$ as the connectivity-conditioned ROI representation. It is used only for effective-connectivity estimation, allowing the edge head to adapt to subject-level variability while preserving ROI-resolved structure.

3.2 DCM Parameter Heads

The effective connectivity head predicts directed coupling using an ordered pairwise interaction between connectivity-conditioned ROI embeddings (see Fig. 1.4):

$$\hat{A}_{ij} = s \cdot \tanh\left(f_A\left(Z_{\mathrm{conn},i}, Z_{\mathrm{conn},j}, g_{\mathrm{subj}}\right)\right),$$

where f_A is a directed pairwise MLP edge head and s is a learnable output scale. Since the ordered pair (i, j) is treated distinctly from (j, i), this parameterization supports asymmetric directed interactions. The bounded nonlinearity stabilizes training while preserving signed edge weights.

The remaining parameters are estimated using heads tailored to their roles in the DCM forward model. Global spectral parameters are estimated from subject-level summaries of the spatial representations, while the ROI-wise parameter c is estimated per region (see Fig. 1.2 and Fig. 1.4):

$$(\hat{a}, \hat{b}) = f_{\mathrm{global}}(Z_{\mathrm{spat}}), \qquad \hat{c}_i = f_c(Z_{\mathrm{spat},i}).$$

Transit time is modeled separately because it directly reflects hemodynamic delay. We first estimate an ROI-wise transit component from temporal and phase representations,

$$\hat{\tau}_i^{\mathrm{pre}} = f_\tau^{\mathrm{pre}}\left(Z_{\mathrm{temp},i}, Z_{\mathrm{phase},i}\right),$$

and then refine it using spatial context:

$$\hat{\tau}_i = \hat{\tau}_i^{\mathrm{pre}} + \eta f_\tau^{\mathrm{ref}}\left(Z_{\mathrm{spat},i}\right).$$

This separates timing-sensitive hemodynamic inference from connectivity prediction, while still allowing spatial context to refine ROI-specific transit estimates.

3.3 Forward-Model Consistency via Analytic CSD

In addition to being interpretable, physics-informed models are governed by equations that can be leveraged to constrain the parameter space and improve learning. BASIL 's model parameters have a biophysical meaning and imply a precise frequency spectrum for the fMRI BOLD time series (more specifically, a precise cross-spectral density, or CSD). This can be derived analytically via the forward model's equations (see Appendix C.3), producing a differentiable CSD forward operator,

$$\hat{\mathcal{G}} = \mathcal{S}_{\mathrm{DCM}}(\hat{\theta}), \qquad \hat{\theta} = \{\hat{A}, \hat{a}, \hat{b}, \hat{c}, \hat{\tau}\},$$

which maps estimated DCM parameters to the model-implied CSD (Fig. 1.5). This analytic operator uses the standard spectral-DCM linearization of the nonlinear hemodynamic equations around the resting state and is distinct from the nonlinear time-domain model used to generate the simulated BOLD sequences. It serves as one regularization term within the composite training objective in Section 3.4. BASIL uses this analytic spectrum to regularize the inverse model in the frequency domain, encouraging the estimated parameters to jointly reproduce the spectral structure implied by the DCM dynamics, rather than being estimated independently. As we will see in the ablation results below, this is a crucial component of the objective function. Finally, one could in principle replace the analytic CSD with the observed BOLD CSD to fine-tune a pretrained BASIL model on data with different acquisition, spectral features, and population characteristics.

3.4 Training Objective

BASIL is trained end-to-end with a composite objective:

$$\mathcal{L} = \mathcal{L}_A + \mathcal{L}_{\mathrm{param}} + \lambda_{\mathrm{CSD}}\mathcal{L}_{\mathrm{CSD}}.$$

The effective-connectivity objective $\mathcal{L}_A$ combines complementary constraints on the estimated directed matrix $\hat{A}$. It includes variance-weighted edge-wise supervision to recover connectivity magnitudes, correlation-based supervision to preserve whole-matrix connectivity structure, sign consistency to recover excitatory/inhibitory polarity on strong edges, a KL term that uses the estimated edge variance to regularize uncertainty against reference posterior variance, a residual term that preserves subject-specific departures from the population mean, and a contrastive term that matches each estimated matrix to its corresponding reference rather than other subjects. Uncertainty estimation is restricted to effective-connectivity edges: for each predicted edge mean $\hat{A}_{ij}$, a separate variance head predicts a positive variance $\hat{V}_{A,ij}$, which is regularized against the corresponding reference posterior variance. For a, b, c, and τ, BASIL currently provides point estimates without associated uncertainty estimates. Uncertainty calibration is not evaluated in the present study. The parameter loss $\mathcal{L}_{\mathrm{param}}$ supervises the remaining (endogenous) neuronal, observation noise, and hemodynamic parameters, including global parameters, ROI-wise parameters, and transit times. The spectral loss

$\mathcal{L}_{\text{CSD}}$ provides frequency-domain forward-model regularization by penalizing parameters whose model-implied spectra deviate from reference spectra.

Together, these objectives train BASIL to infer parameters that are accurate, signed and directed, representation-stable, and consistent with the DCM generative model. Full architectural details, uncertainty modeling, phase feature extraction, loss terms, training steps, and hyperparameters are provided in Appendix B, Algorithm 1, and Tables S2–S3. Source code is available on GitHub.[1]

4 Experiments

We evaluate BASIL using complementary performance metrics: recovery of directed connectivity, neuronal and hemodynamic parameters, identification of sparse and signed structure, and preservation of global network properties including hub ordering, centrality, and modularity.

4.1 Dataset

We simulate resting-state fMRI time series informed by empirical data from 994 subjects in the HCP Young Adult cohort, with 20% of data held out as an independent test set used only for final evaluation. Full detail about the training set generation is provided in Appendix A.

4.2 Baselines

Table 1: **Baselines and the primary hypothesis each tests.**

Method	Purpose / hypothesis tested
rDCM [14]	fMRI-specific baseline: performs Bayesian inversion of a neuronal–hemodynamic model without amortized learning. It is fast due to many simplifying assumptions: 1) the hemodynamic response has a fixed "canonical" shape (not modulated by biophysical parameters), 2) hemodynamic response is identical for all brain regions; 3) connectivity parameters are partially independent and can be inferred in parallel via linear regression.
NGC [32]	Nonlinear predictive-causality baseline: tests whether neural predictors improve directed-interaction recovery under a Granger criterion.
EC-GAN [33]	Generative adversarial baseline: tests whether connectivity can be inferred by learning patterns that reproduce fMRI dynamics.
TCDF [34]	Temporal deep causal discovery baseline: tests whether attention-based dilated CNNs with deconvolution recover nonlinear lagged causes and delays.
CaLLTiF [26]	Constraint-based baseline: tests whether PCMCI-style conditional-independence testing over lagged variables recovers a statistically supported lag-resolved directed graph. For our benchmark, the lag-resolved output is summarized as a binary directed adjacency.
NPI [22]	Perturbational neural baseline: tests whether virtual ROI perturbations of a learned surrogate brain model recover signed directed EC.

We compare BASIL against several methods spanning mechanistic modeling, classical predictive causality, deep learning, and recent neuroscience-specific causal discovery (Table 1; quantitative results in Table 2). As a biophysical reference, we include rDCM, which estimates effective connectivity by inverting a neuronal–hemodynamic model for whole-brain resting-state fMRI analysis [14]. To test predictive formulations, we include NGC, a nonlinear extension of Granger causality that evaluates whether directed interactions can be recovered from nonlinear temporal predictability [32, 23]. We also include deep learning-based causal estimators: TCDF, which combines learned deconvolution with attention-based temporal convolutions to recover nonlinear lagged causal relationships and delays [34], and EC-GAN, which tests whether adversarially learned generative dynamics can infer connectivity patterns that reproduce observed fMRI activity [33]. Finally, we include neuroscience-specific methods: CaLLTiF, which uses PCMCI-style conditional-independence testing over lagged and contemporaneous variables to produce a lag-resolved directed graph with edge-level partial-correlation statistics and p-values [26], and NPI, which estimates effective connectivity through virtual perturbations of a learned surrogate brain model [22]. CaLLTiF therefore does not natively estimate the signed, continuous DCM coupling weights targeted by BASIL, while NPI does not explicitly incorporate biophysical constraints. Notably, only BASIL estimates endogenous neuronal fluctuations, observation noise, and hemodynamic parameters beyond the effective-connectivity graph; the remaining baselines estimate graph structure or coupling weights only. Detailed baseline descriptions, architectures, and hyperparameters are provided in Appendix E. Runtime scalability relative to classical DCM approaches is evaluated in Appendix I.

4.3 Metrics

We evaluate recovery at edge, sign, parameter, and network levels. Edge-level accuracy is measured by Pearson correlation $\rho_A = \rho(A, \hat{A})$, computed over off-diagonal edge weights. Top-k overlap is measured between the strongest estimated and reference edges ranked by absolute value. Signed

[1] https://github.com/basildcmrepo/BASIL-DCM

structure is assessed over off-diagonal entries using sign agreement and sign-confusion matrices. Neuronal, observation noise, and hemodynamic parameter recovery is measured by Pearson correlation for global parameters (a_1, a_2, b_1, b_2) across subjects and ROI-wise parameters (c_i, τ_i) pooled across subjects and ROIs. Network-level validity is evaluated by comparing derived graph properties of A and $\hat{A}$, including incoming/outgoing strength profiles, hub ordering, mean edge weight, modularity, maximum betweenness centrality, and spectral radius.

5 Results

Table 2 summarizes the main comparison to baselines (top block) and BASIL ablations (bottom). The comparison is based on edge-weight recovery (ρ_A), strong-edge identification (Top-40%), and edge-sign classification via true positive and true negative rates (TPR/TNR, conditioned on the true sign). BASIL achieves the strongest recovery of directed connectivity across most metrics, while ablations show that each major architectural and physics-informed component contributes to performance. The low performance of rDCM may seem surprising since we generated the training time series using rDCM-estimated effective connectivity. While this ensures that A is realistic, our generative model intentionally extends beyond several rDCM assumptions (listed in Table 1) by including regional hemodynamic variations and autocorrelated endogenous fluctuations and noise, introducing sources of variability that are not explicitly inferred by rDCM (see Appendix A). Supplementary runtime analysis in Appendix Fig. S7 compares BASIL with existing DCM variants. Runtime analysis showed that sDCM did not complete at $N = 100$, whereas BASIL remained tractable. rDCM was slightly faster but relies on simplifying assumptions and estimates fewer biophysical parameters (Appendix I).

Table 2: **Baseline comparison and ablations.**[†]

Metric \ Method	rDCM	NGC	EC-GAN	TCDF	CaLLTiF	NPI	BASIL
ρ_A	0.069	0.368	0.424	0.610	0.273	0.498	**0.711**
Top-40%	0.419	0.410	0.483	0.559	✗	0.496	**0.626**
True Positive Rate (TPR) %	52.1	✗	58.1	92.4	✗	75.9	**97.6**
True Negative Rate (TNR) %	52.6	✗	53.2	36.8	✗	**64.5**	62.3

Ablation \ Metric	ρ_A	ρ_{a_1}	ρ_{a_2}	ρ_{b_1}	ρ_{b_2}	ρ_c	ρ_τ
A1: GRU temporal encoder	0.653	0.904	**0.931**	0.858	**0.908**	0.913	0.528
A2: Signed GCN spatial encoder	0.528	0.908	0.894	0.874	0.865	0.907	0.375
A3: No spatial encoder	0.330	0.710	0.683	0.674	0.675	0.710	0.263
A4: No CSD loss	0.573	0.815	0.916	0.774	0.894	0.847	0.291
A5: Full BASIL (ours)	**0.711**	**0.926**	0.920	**0.874**	0.900	**0.926**	**0.549**

[†] ρ_A, TPR, and TNR are computed over off-diagonal entries only. NGC estimated only positive edge weights, so sign-based metrics are omitted. For evaluation, CaLLTiF's lag-resolved output is summarized into an unsigned binary graph; non-applicable metrics are omitted, and ρ_A is computed against binarized ground-truth A.

5.1 Parameter recovery and sign structure

Fig. 2A–G reports recovery of directed connectivity, neuronal, observation noise, and hemodynamic parameters on the HCP-derived validation set. BASIL accurately recovers off-diagonal edge weights of the adjacency matrix ($\rho_A = 0.711$) and global neuronal parameters (a_1, a_2, b_1, b_2; $\rho = 0.874$–0.926), while also achieving strong recovery of the mean ROI-wise observation parameter c ($\rho_c = 0.926$). Transit time τ is recovered more moderately ($\rho_\tau = 0.549$), remaining the most challenging parameter due to its weaker identifiability from BOLD dynamics and sensitivity to noise. These HCP-informed simulations assess parameter recovery under the specified generative model, not neuroscientific or clinical validity. Finally, the sign confusion matrix shows that BASIL reliably preserves the polarity of directed interactions (Fig. 2H), suggesting that its estimations are sufficiently sign-consistent to support analyses where excitatory/inhibitory directionality is of interest. A method-wise sign-confusion comparison in Appendix Fig. S1 further shows that BASIL provides the most balanced recovery of signed interactions across baselines.

5.2 Recovery of sparse structure and graph-level organization

Having established direct edge- and matrix-level recovery in Table 2 and Fig. 2 (also see Fig. 4B), Fig. 3 provides an assessment of whether BASIL-estimated matrices preserve graph-level properties for downstream network analysis. Because these summaries are many-to-one, their agreement should not be interpreted as independent evidence of full matrix recovery. Because the CaLLTiF benchmark output is a lag-summarized binary adjacency and does not provide signed, continuous coupling weights, CaLLTiF is not included in the graph-level and weight/sign-based analyses below. For

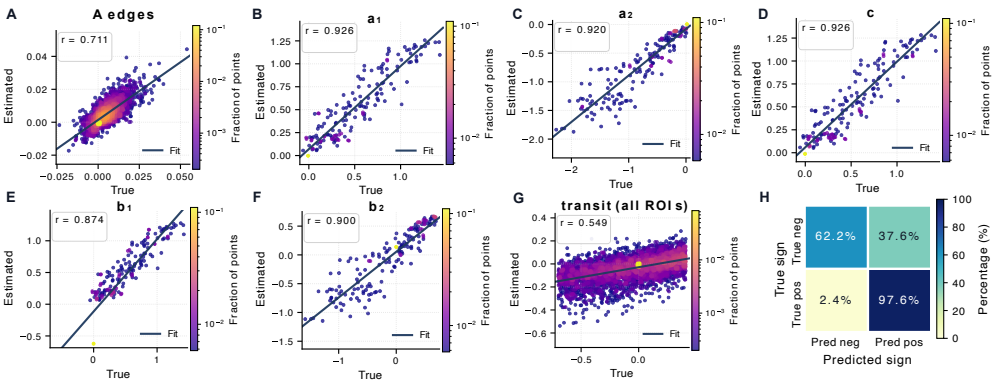


Figure 2: **Model parameter recovery, including connection weight and sign.** (A) Recovery of off-diagonal edge weights in the directed adjacency matrix A. (B–F) Parameters characterizing power-law spectral decay of endogenous neuronal fluctuations (a_1, a_2, b_1, b_2) and mean ROI-wise observation-noise parameter c. (G) ROI-wise transit time. (H) Sign confusion matrix over off-diagonal edges of A.

network-level visualization, ROIs are ordered using the Schaefer-100 atlas [7] and grouped by Yeo-7 resting-state networks: VIS, SOM, DAN, VAN, LIM, FPN, and DMN[2] [35]; the shaded bands in Fig. 3A mark these assignments. At the node level, estimated incoming and outgoing strength profiles—averaged over subjects—closely match the reference across ROIs (Fig. 3A), indicating that BASIL preserves ROI-wise variation in directed connectivity strength across canonical resting-state networks. Consistently, hub-score ordering, based on total incoming plus outgoing absolute strength per ROI, is recovered with high fidelity (Fig. 3C; $r \approx 0.98$), showing the model retains the relative importance of ROIs. At the global level, graph statistics computed from each subject's $\hat{A}$, including mean edge weight, modularity index, maximum betweenness centrality, and spectral radius, agree closely with those from A (Fig. 3B), which can support the use of inferred effective connectivity matrices for downstream network-level analyses. Paired subject-level comparisons, Pearson correlations, and mean absolute errors for the graph metrics in Fig. 3B are shown in Fig. S3.

Extending this graph-level analysis to all baselines confirms that BASIL best matches the reference graph statistics, followed by TCDF, while other methods show larger distributional shifts (Appendix Fig. S2). Furthermore, BASIL achieves the highest sign-detection fidelity across subjects (Fig. 3D), indicating more accurate classification of excitatory versus inhibitory interactions than any baseline. BASIL also attains the strongest top-k overlap over a wide sparsity range (1%–50% of edges; Fig. 3E), demonstrating superior recovery of the *magnitude-based ranking* of the strongest directed connectivity. While individual baselines approach BASIL in one criterion, they fall short on others: BASIL is uniquely strong on both sign fidelity and top-k strength recovery, which is essential to identify the most influential directed interactions.

5.3 Recovery of network organization and spatial motifs

Fig. 4 evaluates whether BASIL preserves mesoscale organization and spatial structure in whole-network effective connectivity. At the group level, estimated connectivity reproduces within-network blocks and prominent between-network motifs of the reference matrix when ROIs are ordered by canonical networks (Fig. 4A). This agreement is consistent across subjects; per-subject edge-wise correlations between true and estimated A are centered at a high value ($\bar{\rho}_A = 0.712$; Fig. 4B, left). Average inter-subject correlations within each functional network block are closely matched between true and estimated A (Fig. 4B, right); For baseline comparisons, see Appendix Fig. S4.

To assess spatial plausibility, we visualize directed seed-connectivity maps on cortical surfaces for representative seeds from each canonical network (Fig. 4C). For each seed ROI (Table S11), we plot its strongest within-network effective connections and compare the reference with the BASIL-estimated map. BASIL estimations (green) recover the dominant spatial motifs and hemispheric patterns seen in the reference maps (pink). Finally, graph structure is preserved: degree distributions for $\hat{A}$ closely match those of A, and top-degree ROIs show similar cortical locations between true and estimated maps (Fig. 4D-E), indicating preservation of sparsity structure and hub localization. For

[2]VIS: Visual; SOM: Somatomotor; DAN: Dorsal Attention Network; VAN: Ventral Attention Network; LIM: Limbic; FPN: Frontoparietal Network; DMN: Default Mode Network.

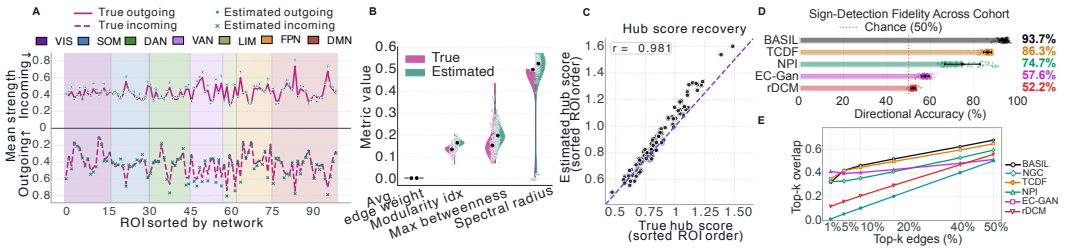


Figure 3: **Graph-level recovery and baseline comparison.** (A) Mean incoming and outgoing ROI strength profiles averaged over subjects, with ROIs ordered by resting-state networks. (B) Agreement between A and $\hat{A}$ across global graph metrics; violins show subject-level distributions, and black dots indicate mean values. (C) Hub-score recovery, measuring preservation of high-influence ROI ordering across all subjects. (D) Sign-detection fidelity across subjects and baselines; bars show cohort means, points show subject-level accuracies, error bars show standard deviation, and the dashed line marks chance (50%). (E) Top-k overlap between the true and estimated sets of strongest directed edges, plotted as a function of the retained edge fraction (ranking edges by $|A_{ij}|$).

completeness, we report supplementary results on a smaller Default Mode Network (DMN) subset, studied in DCM and related effective-connectivity baselines (Appendix G, Figs. S5,S6).

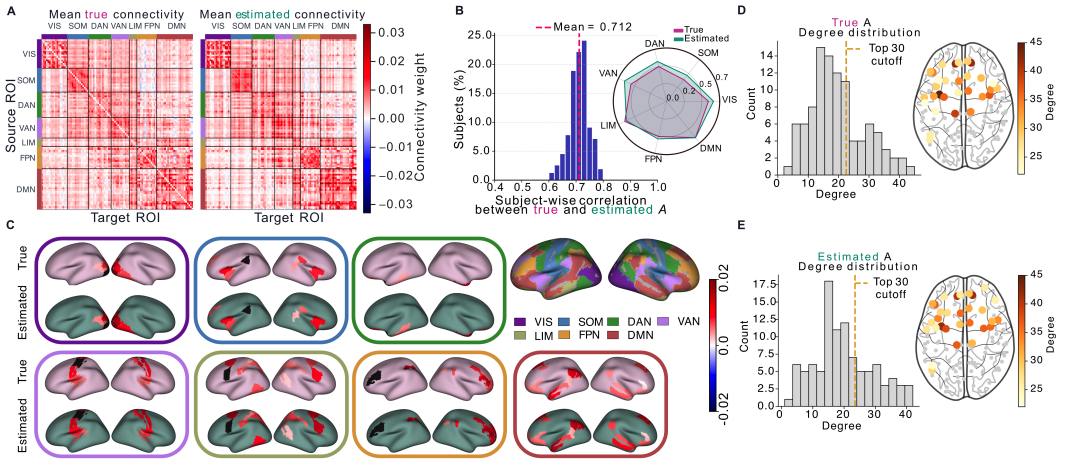


Figure 4: **Network recovery analysis. (A)** Group-averaged true (A; left) and BASIL -estimated A (right) effective connectivity matrices, with ROIs ordered by canonical resting-state networks and shown on the same color scale. Entries denote directed coupling weights A_{ij} (source $j \to i$); red/blue indicate positive/negative influences and intensity indicates magnitude. The close correspondence illustrates preservation of within- and between-network connectivity structure. **(B)** left, histogram of per-subject Pearson correlation between true and estimated A (ρ_A), with mean indicated (0.712); right, true (pink) and estimated (green) average pairwise correlation between subjects, computed within each functional network block, showing which network distributions are most faithfully recovered. **(C)** Directed seed-connectivity maps on the cortical surface for representative seed ROIs from each resting-state network. For each seed, we visualize the strongest within-network effective connections between that seed and other ROIs, comparing the reference map (top) with the BASIL estimate (bottom). Seed locations are indicated by black markers when visible on the displayed cortical view. Maps are thresholded to the strongest within-network connections for visual comparability. **(D-E)** Hub location. For each matrix, histograms show ROI degree distributions with dashed lines marking the top-30 degree cutoff; adjacent brain maps show the corresponding high-degree ROIs, demonstrating preservation of sparsity structure and hub localization.

5.4 Empirical validation and transfer.

To complement simulation-based recovery, we evaluated BASIL on real HCP BOLD from a 24-ROI DMN, where spectral DCM is computationally feasible. BASIL estimates agreed with spectral DCM across effective connectivity and biophysical parameters ($r = 0.392$–0.847; $r_A = 0.704$). To assess stability across independent data segments, we tested split-half reliability using non-

overlapping BOLD segments. Estimates of A were highly reproducible ($r = 0.909$, ICC$(3,1) = 0.909$, MAE $= 0.0040$, sign agreement $= 99.4\%$). Parameters inferred from one half also predicted the empirical CSD of the other ($r = 0.905 \pm 0.055$), substantially exceeding shuffled-subject pairings (0.648 ± 0.005; permutation $p = 0.0099$).

To assess downstream utility, we used BASIL-estimated A alone to predict gender in 200 held-out subjects. Under matched repeated stratified five-fold cross-validation, BASIL achieved balanced accuracy 0.844 and ROC-AUC 0.896, compared with 0.634 and 0.694 for rDCM and 0.747 and 0.747 for functional connectivity, respectively.

Finally, using PsiConnect-informed simulations [36] ($N = 62$), fine-tuning the HCP-pretrained model achieved $\rho_A = 0.598$, compared with 0.643 for dataset-specific training from scratch. This provides initial evidence of cross-dataset adaptation across independently informed parameter distributions within the same DCM simulation family. Full details are in Appendix H, Tables S6-S10.

5.5 Ablations

The ablation block in Table 2 isolates the contribution of each major component of BASIL . Replacing the selective state-space temporal encoder with a bidirectional GRU (A1) consistently reduces connectivity recovery and strong-edge overlap, indicating that long-context sequence modeling is critical for disentangling delayed hemodynamics from directed connectivity.

On the spatial side, replacing the Transformer with a signed directed graph convolutional network (A2) also degrades performance. This ablation uses a learnable, fully connected signed directed adjacency for message passing, allowing weighted asymmetric interactions but using a subject-shared propagation structure rather than input-adaptive attention. Its lower performance suggests that flexible, subject-specific cross-regional interactions are important for accurately recovering effective connectivity from BOLD dynamics.

Removing the spatial encoder entirely (A3) further degrades performance across all metrics, confirming that explicit inter-regional interaction modeling is necessary to infer directed connectivity. Finally, removing the spectral-consistency (CSD) regularizer (A4) weakens recovery of frequency-sensitive neuronal and hemodynamic parameters, leading to less faithful dynamical structure.

Overall, the full model (A5) gives the best trade-off across connectivity recovery, sparse-structure identification, and biophysical parameter estimation. Further details are in Appendix F.

6 Conclusion

We introduced BASIL , a physics-informed amortized framework for scalable cortex-wide estimation of DCM-style effective connectivity and biophysical parameters from resting-state fMRI. Combining a linear-time selective state-space temporal encoder with an ROI-wise Transformer, BASIL infers subject-specific directed connectivity together with neuronal, observation-noise, and hemodynamic parameters in a single forward pass, extending beyond scalable baselines that estimate connectivity alone. Rather than iterative per-subject optimization, the model is trained by supervised amortization on HCP-informed simulated resting-state BOLD, using rDCM-derived parameter estimates as targets and regularized with a differentiable spectral-consistency objective. Across extensive evaluations, BASIL achieves strong recovery of directed connectivity, improves identification of strong edges and signed structure, preserves global network organization, including hub score and location, and remains compatible with modern parcellation sizes and cohort-scale analyses. We further provide an initial demonstration of cross-dataset transfer and pretrained-model reuse through lightweight adaptation to the independent PsiConnect cohort.

Limitations. Several limitations and broader-impact considerations remain. Because supervision is derived from rDCM estimates rather than invasive ground truth, BASIL inherits assumptions and biases of the teacher model. Parameter identifiability is conditional on the assumed DCM forward model, fixed hemodynamic parameters, and specified priors. Effective connectivity, hemodynamic delay, endogenous fluctuations, and observation noise can have overlapping effects on resting-state BOLD spectra, with transit time remaining the least recoverable parameter. Additional hemodynamic parameters (e.g., signal decay κ and neurovascular coupling ϵ) are not yet modeled. While scalable effective-connectivity estimation could support large-cohort neuroscience and disease-related network studies, inferred connectivity should be treated as a research tool rather than a clinical decision system without further validation, especially across acquisition sites and populations. The PsiConnect experiment remains within the same DCM simulation family; future work will test broader transfer across acquisitions and generative assumptions, assess edge-variance calibration, and extend uncertainty estimation to the remaining biophysical parameters.

Acknowledgments and disclosures

A.R. received funding from the Australian Research Council (Refs: DP200100757, FT250100563, and DP260104251), the Australian National Health and Medical Research Council (Investigator Grant Ref: 1194910). A.R. and M.K. are also funded by the Australian Office of National Intelligence (ONI) through their National Intelligence and Discovery Grant scheme (Ref: NI230100153). A.R. is also funded by Wellcome Leap's 'Untangling Addiction' program grant. A.R. is a CIFAR Global Scholar in the Brain, Mind & Consciousness Program and acknowledges support through CIFAR Next-Generation Initiatives (GS22-009) and Catalyst Grants (CF-0224 and CF-0491).

The authors declare no competing interests.

References

[1] Adeel Razi and Karl J. Friston. The Connected Brain: Causality, models, and intrinsic dynamics. *IEEE Signal Processing Magazine*, 33(3):14–35, May 2016.

[2] Karl J. Friston. Functional and Effective Connectivity: A Review. *Brain Connectivity*, 1(1):13–36, January 2011.

[3] Richard B Buxton, Eric C Wong, and Lawrence R Frank. Dynamics of blood flow and oxygenation changes during brain activation: the balloon model. *Magnetic resonance in medicine*, 39(6):855–864, 1998.

[4] Karl J. Friston, Andrea Mechelli, Robert Turner, and Cathy J. Price. Nonlinear responses in fmri: the balloon model, volterra kernels, and other hemodynamics. *NeuroImage*, 12(4):466–477, 2000.

[5] Klaas Enno Stephan, Nikolaus Weiskopf, Peter M Drysdale, Peter A Robinson, and Karl J Friston. Comparing hemodynamic models with dcm. *Neuroimage*, 38(3):387–401, 2007.

[6] Karl J. Friston, Lee Harrison, and William Penny. Dynamic causal modelling. *NeuroImage*, 19(4):1273–1302, 2003.

[7] Alexander Schaefer, Ru Kong, Evan M Gordon, Timothy O Laumann, Xi-Nian Zuo, Avram J Holmes, Simon B Eickhoff, and BT Thomas Yeo. Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity mri. *Cerebral cortex*, 28(9):3095–3114, 2018.

[8] David C Van Essen, Kamil Ugurbil, Edward Auerbach, Deanna Barch, Timothy EJ Behrens, Richard Bucholz, Acer Chang, Liyong Chen, Maurizio Corbetta, Sandra W Curtiss, et al. The human connectome project: a data acquisition perspective. *Neuroimage*, 62(4):2222–2231, 2012.

[9] HCP Wu-Minn. 1200 subjects data release reference manual. *URL https://www. humanconnectome. org*, 565(2), 2017.

[10] Cathie Sudlow, John Gallacher, Naomi Allen, Valerie Beral, Paul Burton, John Danesh, Paul Downey, Paul Elliott, Jane Green, Martin Landray, et al. Uk biobank: an open access resource for identifying the causes of a wide range of complex diseases of middle and old age. *PLoS medicine*, 12(3):e1001779, 2015.

[11] Kyle Cranmer, Johann Brehmer, and Gilles Louppe. The frontier of simulation-based inference. *Proceedings of the National Academy of Sciences*, 117(48):30055–30062, 2020.

[12] Jan-Matthis Lueckmann, Jan Boelts, David Greenberg, Pedro Goncalves, and Jakob Macke. Benchmarking simulation-based inference. In *International conference on artificial intelligence and statistics*, pages 343–351. PMLR, 2021.

[13] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

[14] Stefan Frässle, Samuel J Harrison, Jakob Heinzle, Brett A Clementz, Carol A Tamminga, John A Sweeney, Elliot S Gershon, Matcheri S Keshavan, Godfrey D Pearlson, Albert Powers, et al. Regression dynamic causal modeling for resting-state fmri. *Human brain mapping*, 42(7):2159–2180, 2021.

[15] Karl J Friston, Joshua Kahan, Bharat Biswal, and Adeel Razi. A dcm for resting state fmri. *Neuroimage*, 94:396–407, 2014.

[16] A. Razi, J. Kahan, G. Rees, and K. J. Friston. Construct validation of a DCM for resting state fMRI. *Neuroimage*, 106:1–14, February 2015.

[17] Leonardo Novelli, Karl Friston, and Adeel Razi. Spectral dynamic causal modeling: A didactic introduction and its relationship with functional connectivity. *Network Neuroscience*, 8(1):178–202, 2024.

[18] Stefan Frässle, Ekaterina I. Lomakina, Adeel Razi, Karl J. Friston, Joachim M. Buhmann, and Klaas E. Stephan. Regression DCM for fMRI. *NeuroImage*, 155:406–421, July 2017.

[19] Kamen A Tsvetanov, Richard NA Henson, and James B Rowe. Separating vascular and neuronal effects of age on fmri bold signals. *Philosophical Transactions of the Royal Society B*, 376(1815):20190631, 2021.

[20] Costantino Iadecola. Neurovascular regulation in the normal brain and in alzheimer's disease. *Nature Reviews Neuroscience*, 5(5):347–360, 2004.

[21] Adeel Razi, Mohamed L. Seghier, Yuan Zhou, Peter McColgan, Peter Zeidman, Hae-Jeong Park, Olaf Sporns, Geraint Rees, and Karl J. Friston. Large-scale DCMs for resting-state fMRI. *Network neuroscience (Cambridge, Mass.)*, 1(3):222–241, 2017.

[22] Zixiang Luo, Kaining Peng, Zhichao Liang, Shengyuan Cai, Chenyu Xu, Dan Li, Yu Hu, Changsong Zhou, and Quanying Liu. Mapping effective connectivity by virtually perturbing a surrogate brain. *Nature Methods*, 22(6):1376–1385, 2025.

[23] Clive WJ Granger. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: journal of the Econometric Society*, pages 424–438, 1969.

[24] Jakob Runge, Peer Nowack, Marlene Kretschmer, Seth Flaxman, and Dino Sejdinovic. Detecting and quantifying causal associations in large nonlinear time series datasets. *Science advances*, 5(11):eaau4996, 2019.

[25] Leonardo Novelli, Patricia Wollstadt, Pedro Mediano, Michael Wibral, and Joseph T Lizier. Large-scale directed network inference with multivariate transfer entropy and hierarchical statistical testing. *Network Neuroscience*, 3(3):827–847, 2019.

[26] Fahimeh Arab, AmirEmad Ghassami, Hamidreza Jamalabadi, Megan AK Peters, and Erfan Nozari. Whole-brain causal discovery using fmri. *Network Neuroscience*, 9(1):392, 2025.

[27] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[28] Moein Khajehnejad, Forough Habibollahi, Devon Stoliker, and Adeel Razi. Brainsymphony: A parameter-efficient multimodal foundation model for brain dynamics with limited data, 2026.

[29] Yuqing Xia, Fahimeh Arab, Urmi Saha, Benjamin Sipes, Grayson Gooden, Minghan Chen, and Ashish Raj. Mambaxbrain: A multi-task neural framework linking brain functional dynamics to individual fingerprints, cognitive and disease states. *bioRxiv*, pages 2026–02, 2026.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

[31] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.

[32] Alex Tank, Ian Covert, Nicholas Foti, Ali Shojaie, and Emily B Fox. Neural granger causality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(8):4267–4279, 2021.

[33] Jinduo Liu, Junzhong Ji, Guangxu Xun, Liuyi Yao, Mengdi Huai, and Aidong Zhang. Ec-gan: Inferring brain effective connectivity via generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 4852–4859, 2020.

[34] Meike Nauta, Doina Bucur, and Christin Seifert. Causal discovery with attention-based convolutional neural networks. *Machine Learning and Knowledge Extraction*, 1(1):19, 2019.

[35] BT Thomas Yeo, Fenna M Krienen, Jorge Sepulcre, Mert R Sabuncu, Danial Lashkari, Marisa Hollinshead, Joshua L Roffman, Jordan W Smoller, Lilla Zöllei, Jonathan R Polimeni, et al. The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *Journal of neurophysiology*, 2011.

[36] Leonardo Novelli, Devon Stoliker, Tamrin Barta, Matthew D Greaves, Sidhant Chopra, James Jackson, Jessica Kwee, James C Pang, Martin L Williams, and Adeel Razi. Psiconnect: Multimodal neuroimaging of context-dependent brain and behaviour dynamics under psilocybin. *Scientific Data*, 2026.

[37] N.J. Kasdin. Discrete simulation of colored noise and stochastic processes and 1/f^α power law noise generation. *Proceedings of the IEEE*, 83(5):802–827, 1995.

[38] Takayuki Obata, Thomas T Liu, Karla L Miller, Wen-Ming Luh, Eric C Wong, Lawrence R Frank, and Richard B Buxton. Discrepancies between bold and flow dynamics in primary and supplementary motor areas: application of the balloon model to the interpretation of bold transients. *NeuroImage*, 21(1):144–153, 2004.

[39] Robert L Grubb Jr, Marcus E Raichle, John O Eichling, and Michel M Ter-Pogossian. The effects of changes in paco2 cerebral blood volume, blood flow, and vascular mean transit time. *Stroke*, 5(5):630–639, 1974.

[40] Leonardo Novelli, Lionel Barnett, Anil K Seth, and Adeel Razi. Minimum-phase property of the hemodynamic response function, and implications for granger causality in fmri. *Human Brain Mapping*, 46(10):e70285, 2025.

[41] Martin Havlicek, Alard Roebroeck, Karl Friston, Anna Gardumi, Dimo Ivanov, and Kamil Uludag. Physiologically informed dynamic causal modeling of fMRI data. *NeuroImage*, 122:355–372, 2015.

[42] Kâmil Uludağ, Bernd Müller-Bierl, and Kâmil Uğurbil. An integrative model for neuronal activity-induced signal changes for gradient and spin echo functional imaging. *NeuroImage*, 48(1):150–165, October 2009.

[43] Baojuan Li, Jean Daunizeau, Klaas E Stephan, Will Penny, Dewen Hu, and Karl Friston. Generalised filtering and stochastic dcm for fmri. *neuroimage*, 58(2):442–457, 2011.

A Generative model details and priors

A.1 HCP Preprocessing Pipeline

We used ICA-FIX-denoised resting-state fMRI data from the HCP Young Adult cohort. The HCP minimal preprocessing pipeline included gradient-nonlinearity correction, rigid-body motion correction, susceptibility-distortion correction, registration to the individual anatomical image and standard space, intensity normalisation, and p. Structured artefactual components related to motion, physiological fluctuations, and scanner effects were subsequently removed using ICA-FIX.

No additional global signal regression was applied. We now state this decision explicitly, since global signal regression can modify individual connectivity estimates and, in some cases, their signs. Accordingly, the negative connections in the rDCM-derived reference matrices were not introduced by an additional global-signal-regression step. We have also specified the temporal filtering and detrending applied to the released HCP data, whether any additional motion censoring or nuisance regression was performed, and the standardisation applied before rDCM inversion.

For regional time-series extraction, the Schaefer-100 atlas was used. For each subject, the BOLD time series for each parcel was obtained by extracting the first eigenvariate of all voxel time series within that parcel. We have additionally reported the exact HCP release, resting-state run or runs, and input-file convention used in the analysis. Separately, we agree that the Schaefer-100 atlas provides cortical coverage only. We have therefore replaced claims of "whole-brain" analysis with "cortex-wide" analysis throughout the manuscript and now state explicitly in Section 4.1 that subcortical and cerebellar regions were not included.

A.2 Neuronal State Dynamics

In the following sections, we distinguish latent neuronal states (x) from observed BOLD signals (y), following the standard spectral DCM formulation [15, 17].

Let $x(t) \in \mathbb{R}^N$ denote latent neuronal activity across N brain regions. We assume linear stochastic dynamics

$$\dot{x}(t) = Ax(t) + u(t), \tag{1}$$

where $A \in \mathbb{R}^{N\times N}$ is the effective connectivity matrix and $u(t)$ is a stochastic term modeling endogenous fluctuations in neuronal activity. The real parts of the eigenvalues of A must be negative for stability, which is typically enforced through priors on self-inhibition and coupling strength [15]. Endogenous fluctuations are modeled as colored noise processes that are independent across regions but share the same power spectrum $\mathcal{G}_u(\omega) = a_1/\omega^{a_2}$, a power-law functional form modeling long-memory dynamics observed in resting-state signals [15, 17].

Sampling endogenous fluctuations We sampled these $u(t)$ time series using the discrete power-law formulation from [37], where a white Gaussian input sequence is filtered by the transfer function

$$H(z) = (1 - z^{-1})^{-a_2/2}$$

producing a process with power spectral density proportional to ω^{-a_2}. The finite impulse-response coefficients were computed recursively as

$$h_0 = 1, h_k = \frac{(a_2/2 + k - 1)}{k}\, h_{k-1}.$$

For each realization, zero-mean Gaussian noise samples were drawn with variance $a_1\, \Delta t^{a_2}$, where Δt denotes the sampling interval and a_1 sets the amplitude of the final power spectrum, $\mathcal{G}_u(\omega) = a_1/\omega^{a_2}$.

A.3 Hemodynamic Observation Model

Observed BOLD time series $y(t) \in \mathbb{R}^N$ are generated by the hemodynamic model proposed by Stephan et al. [5]. This builds on the Balloon-Windkessel equations and previous models [3, 4, 38], introducing four intermediate states per region: vasodilatory signal (s), blood flow (f), blood volume (v), and deoxyhemoglobin content (q). Specifically, for each regional latent neuronal activity $x_i(t)$ from Eq. (1), with $i = 1, \ldots, N$, the model involves four hemodynamic state equations and one

BOLD observation equation:

$$\dot{s}_i(t) = -ks_i(t) - \gamma(f_i(t) - 1) + x_i(t) \tag{2a}$$

$$\dot{f}_i(t) = s_i(t) \tag{2b}$$

$$\dot{v}_i(t) = \left(f_i(t) - v_i(t)^{1/\alpha}\right)/\tau \tag{2c}$$

$$\dot{f}_i(t) = \left(\frac{f_i(t)\left(1 - (1 - E_0)^{1/f_i(t)}\right)}{E_0} - \frac{q_i(t)v_i(t)^{1/\alpha}}{v_i(t)}\right)/\tau \tag{2d}$$

$$y_i(t) = V_0\left(k_1(1 - q_i(t)) + k_2\left(1 - q_i(t)/v_i(t)\right) + k_3(1 - v_i(t))\right) + e_i(t) \tag{2e}$$

This model consists of three parts, visually summarized in [5, Fig. 1]. The first part describes the link between neural activity and regional cerebral blood flow using linear differential equations modelling a dampened oscillator (neurovascular state Eqs. (2a–2b)). Specifically, the neural activity elicits an exponentially decaying vasodilatory signal s that drives blood flow f. In turn, the flow regulates s via a feedback mechanism [4]. The parameter k in Eq. (2a) is the signal decay rate, and γ is the feedback regulation rate (Table S1). Together, these parameters regulate the system to produce damped oscillations with a frequency of approximately 0.09 Hz. The flow f is normalized with respect to the resting flow, so the feedback regulation term $(f - 1)$ in Eq. (2a) becomes zero during rest.

The second part is the hemodynamic model in Eqs. (2c–2d), which builds on the Balloon-Windkessel equations and previous models [3, 4, 38]. It describes the behavior of the post-capillary venous compartment, modeled as an expandable balloon fed by the output of the capillary bed. The model consists of coupled equations representing mass balance for blood volume (v) and deoxyhemoglobin content (q) as they pass through the venous balloon, both normalized to their resting values. The blood volume change rate (v) is modeled as the difference between blood inflow and outflow in Eq. (2c). The blood outflow function is based on Grubb et al., 1974 [39]'s finding that the steady-state relationship between cerebral blood flow and volume follows the power law $v^{1/\alpha}$, where α is the balloon stiffness (see Table S1). Similarly, the rate of change of deoxyhemoglobin (q) is modeled in Eq. (2d) as the difference between the rate of delivery of deoxyhemoglobin into the venous balloon and its clearance rate from the tissue. The oxygen extraction fraction is modeled as a fixed nonlinear function of blood inflow, where E_0 is the net oxygen extraction at rest [3]. This reflects the physiological observation that a large increase in blood flow is required to support a small increase in oxygen metabolism. Both equations are scaled by the mean transit time of blood through the veins.

The third part of the model is the output or BOLD signal change in Eq. (2e), which links venous volume (v) and deoxyhemoglobin content (q) to BOLD signal change [3]. This equation consists of four terms that respectively describe: 1) the extravascular signal as a function of deoxyhemoglobin content, 2) the intravascular signal, given by the ratio of deoxyhemoglobin to blood volume, 3) the volume-weighted balance between extravascular and intravascular signals, and 4) the observation noise $e_i(t)$. The latter is modeled as a colored noise process with power spectrum $\mathcal{G}_e(\omega) = b_1/\omega^{b_2} + c_i/\omega^{b_2}$, which represent a global noise component and region-specific components, respectively. V_0 represents the resting blood volume fraction. The parameters k_1, k_2, and k_3 are dimensionless constants that reflect baseline physiological properties of brain tissue and acquisition parameters of the gradient echo sequence. See Table S1 and [40] for further details and analysis of the model parameters and properties.

Interpretation and potential utility of the inferred parameters. The parameters a_1 and a_2 characterize the amplitude and spectral exponent of endogenous neuronal fluctuations, respectively, and therefore describe the scale and frequency dependence of the latent neuronal drive. They can help distinguish variation in intrinsic neuronal dynamics from variation in directed coupling. The parameters b_1 and b_2 characterize the amplitude and spectral exponent of the global observation noise, while c_i represents the ROI-specific observation-noise component. These parameters help distinguish effective-connectivity differences from BOLD variability associated with measurement or regional observation processes. Finally, the ROI-wise transit time τ_i characterizes regional hemodynamic delay. Accounting for regional differences in hemodynamic timing reduces the risk that vascular-response delays are incorrectly attributed to directed neuronal interactions. Together, these quantities separate complementary contributors to resting-state BOLD dynamics: directed neuronal coupling, endogenous neuronal fluctuations, observation noise, and regional hemodynamic variability. They may therefore provide mechanistically interpretable features for future studies of

Table S1: Parameter interpretation and default values in [5]. The values in brackets are the Bayesian priors for the free parameters used in SPM12. The BOLD model parameter values are for a gradient-echo sequence at 1.5 Tesla.

Parameters and variables	Symbol	Value (Stephan et al. [5])
Variables		
Vasodilatory signal	s	
Blood flow	f	
Venous blood volume	v	
Venous deoxyhemoglobin level	q	
BOLD signal change (%)	y	
Latent neuronal activity	x	
Neurovascular coupling parameters		
Decay of vasodilatory signal	k	(0.64)
Feedback regulation	γ	0.32
Hemodynamic model parameters		
Mean transit time	τ	(2)
Balloon stiffness (Grubb's exponent)	α	0.32
Oxygen extraction factor at rest	E_0	0.4
BOLD model parameters		
Field strength (Tesla)	B_0	1.5
Echo time (s)	T_{E}	0.04
Venous blood volume fraction at rest	V_0	0.04
Frequency offset at vessel surface for deoxygenated blood	θ_0	40.3
Ratio of intra to extravascular signal	ϵ	(1.28)
Sensitivity of intravascular signal relaxation rate to oxygen saturation	r_0	25
1st BOLD parameter	k_1	$4.3\theta_0 E_0 T_{\mathrm{E}} = 2.77$
2nd BOLD parameter	k_2	$\epsilon r_0 E_0 T_{\mathrm{E}} = 0.4\epsilon$
3rd BOLD parameter	k_3	$1 - \epsilon$

aging, pharmacological effects, and neurological or psychiatric conditions. The empirical analyses in Appendix H provide an initial real-BOLD assessment of parameter agreement, split-half reliability, and subject-level information in the 24-ROI DMN, without establishing clinical biomarker validity.

A.4 Parcellation geometry

All simulations are defined on the Schaefer-100 parcellation [7]. Each simulated subject shares the same node identities and centroid coordinates $C = \{c_i\}_{i=1}^{N}$ with $N = 100$. These coordinates define geometry-conditioned connectivity priors, while learnable ROI identity embeddings provide positional information to the spatial encoder during training and evaluation. This ensures that simulation and HCP preprocessing use identical node semantics.

A.5 Ground-truth for training and validation

In the absence of a known parametric form describing the family of effective connectivity matrices, we use an existing scalable DCM variant (rDCM) to infer realistic A matrices using empirical resting-state fMRI data from the HCP. Since rDCM doesn't infer the other biophysical parameters of interest for our generative model, we randomly sample these from known realistic ranges [41, 42]. Specifically, region-dependent hemodynamic transit times are independently sampled within the range $\tau_i \in [1 - 3], \forall i = 1, \ldots, N$. The parameters describing the power spectrum of endogenous neuronal fluctuations and observation noise are sampled to yield $1/f$-like spectra within physiologically plausible ranges ($a_1, b_1, c \in [0.025, 0.033], a_2, b_2 \in [0.1 - 1]$). While it is possible to vary c_i across regions, we set $c_i = c, \forall i = 1, \ldots, N$ for simplicity. This retains region-specific noise modelling, but assumes that the observation noise has the same power spectrum across regions. Together, the A matrices and the other randomly sampled parameters form the ground-truth for training and validating BASIL .

A.6 Time series generation for training and validation

We integrate the neuronal Eq. (1) and the hemodynamic Eqs. (2a–2d), and then compute the simulated BOLD time series via Eq. (2e), where the observation noise $e(t)$ is sampled using the same algorithm described in Appendix A for the endogenous fluctuations. To match the empirical HCP data, we generate $T = 1200$ samples at TR $= 0.72$s.

B Additional Architecture Details

This section expands the architectural components introduced in the main text. We use the following latent representations throughout:

$$Z_{\text{temp}} = f_{\text{temp}}(Y), \qquad Z_{\text{phase}} = f_{\text{phase}}(\mathcal{H}(Y)), \qquad Z_{\text{spat}} = f_{\text{spatial}}(Z_{\text{temp}}),$$

where Z_{temp} denotes Mamba-based temporal ROI embeddings, Z_{phase} denotes phase-sensitive embeddings extracted from the Hilbert-phase representation of the BOLD signal, and Z_{spat} denotes spatially contextualized ROI embeddings.

Temporal encoder f_{temp}. For each ROI time series $Y_i \in \mathbb{R}^T$, the temporal encoder produces hidden states

$$h_{i,t} = f_{\text{seq}}(Y_i)_t, \qquad t = 1, \ldots, T,$$

where f_{seq} is implemented using a selective state-space model and $h_{i,t}$ captures the temporal context up to time t. To obtain a fixed-length ROI representation, we apply attention pooling:

$$Z_{\text{temp},i} = \sum_{t=1}^{T} \alpha_{i,t} h_{i,t}, \qquad \alpha_{i,t} = \frac{\exp(w^\top h_{i,t})}{\sum_{t'=1}^{T} \exp(w^\top h_{i,t'})}.$$

Thus $Z_{\text{temp}} \in \mathbb{R}^{B \times R \times d_{\text{time}}}$ summarizes local ROI dynamics while preserving long-context temporal information. This mechanism allows the model to focus on informative time points while suppressing noise. From a DCM perspective, this stage can be interpreted as learning a compact representation of latent neuronal dynamics from observed BOLD signals, prior to modeling interactions between regions.

Phase encoder f_{phase}. In parallel, we extract phase-sensitive timing features from the analytic signal. For each ROI time series Y_i, the Hilbert transform is

$$\mathcal{H}(Y_i)(t) = \frac{1}{\pi}\,\text{p.v.} \int \frac{Y_i(\tau)}{t - \tau} d\tau,$$

where p.v. denotes the Cauchy principal value, used to handle the singularity at $\tau = t$ and the analytic signal is

$$\tilde{Y}_i(t) = Y_i(t) + j\mathcal{H}(Y_i)(t).$$

The instantaneous phase is

$$\phi_i(t) = \angle \tilde{Y}_i(t).$$

The phase sequence is processed by:

$$Z_{\text{phase},i} = f_{\text{phase}}(\phi_i),$$

where f_{phase} is a multiscale temporal convolutional neural network (CNN). In implementation, f_{phase} first extracts multiscale temporal features from the instantaneous phase sequence and then applies adaptive temporal average pooling, yielding a static ROI-wise phase embedding. The use of multiple temporal receptive fields enables the model to capture phase dynamics at different temporal scales, including both short-term fluctuations and longer-range oscillatory patterns. The resulting phase embeddings are combined with the temporal representations derived from the raw BOLD signal.

From a modeling perspective, these phase features provide additional information about the temporal structure of neuronal activity, which is particularly relevant for estimating timing-related parameters such as the hemodynamic transit time τ.

Spatial encoder f_{spatial}. The temporal ROI embeddings are first projected from d_{time} to the spatial embedding dimension d_{node}:

$$\widetilde{Z}_{\text{temp}} = Z_{\text{temp}} W_{\text{proj}}, \qquad \widetilde{Z}_{\text{temp}} \in \mathbb{R}^{B \times R \times d_{\text{node}}},$$

where W_{proj} is a learned linear projection. The projected ROI embeddings are then contextualized by the spatial Transformer:

$$Z_{\text{spat}} = f_{\text{spatial}}\left(\widetilde{Z}_{\text{temp}}\right).$$

For a single self-attention layer, this can be written as

$$Q = \widetilde{Z}_{\text{temp}} W_Q, \qquad K = \widetilde{Z}_{\text{temp}} W_K, \qquad V = \widetilde{Z}_{\text{temp}} W_V,$$

$$Z_{\text{spat}} = \text{Attn}\left(\widetilde{Z}_{\text{temp}}\right) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_{\text{node}}}}\right) V.$$

The output $Z_{\text{spat}} \in \mathbb{R}^{B \times R \times d_{\text{node}}}$ contains ROI embeddings contextualized by all other regions.

Subject-level modulation and connectivity representation. To enable subject-specific adaptation, we compute a global subject embedding by mean-pooling the spatially contextualized ROI embeddings and passing the result through a small multilayer perceptron:

$$g_{\text{subj}} = f_{\text{subj}}\left(\frac{1}{R}\sum_{i=1}^{R} Z_{\text{spat},i}\right),$$

where f_{subj} denotes the subject-summary MLP.

This embedding modulates ROI representations via feature-wise linear modulation (FiLM):

$$Z_{\text{conn}} = f_{\text{FiLM}}(Z_{\text{spat}}, g_{\text{subj}}) = \gamma(g_{\text{subj}}) \odot Z_{\text{spat}} + \beta(g_{\text{subj}}),$$

where $\gamma(\cdot)$ and $\beta(\cdot)$ are learned affine projections from the subject embedding to feature-wise scale and shift parameters.

The resulting Z_{conn} is used only for effective-connectivity estimation.

Parameter heads f_{head}. The parameter head collection is

$$f_{\text{head}} = \left\{f_{\text{subj}}, f_{\text{FiLM}}, f_{\text{edge}}, f_{\text{global}}, f_c, f_\tau^{\text{pre}}, f_\tau^{\text{ref}}\right\}.$$

Here f_{subj} is a subject-summary MLP, f_{FiLM} is a FiLM modulation module that produces connectivity-conditioned ROI embeddings Z_{conn}, and $f_{\text{edge}} = \{f_A, f_V\}$ is a directed pairwise edge head for effective-connectivity estimation and uncertainty estimation. The remaining heads are MLPs that predict global, ROI-wise, and transit-time parameters.

For each ordered ROI pair (i, j), the effective-connectivity head f_{edge} predicts both a bounded posterior mean and an edge-wise log-variance. We write the mean prediction as

$$\hat{A}_{ij} = s \cdot \tanh\left(f_A\left(Z_{\text{conn},i}, Z_{\text{conn},j}, g_{\text{subj}}\right)\right),$$

where s is a learnable output scale. The corresponding variance is estimated by a separate variance branch:

$$\ell_{ij} = f_V\left(Z_{\text{conn},i}, Z_{\text{conn},j}, g_{\text{subj}}\right), \qquad \hat{V}_{A,ij} = \text{softplus}(\ell_{ij}) + \epsilon_V.$$

Both f_A and f_V are implemented as MLPs over directed pairwise features derived from $(Z_{\text{conn},i}, Z_{\text{conn},j}, g_{\text{subj}})$. Because the edge head receives the ordered pair (i, j), the representation for $i \to j$ differs from that for $j \to i$, allowing asymmetric directed connectivity. The bounded nonlinearity stabilizes the mean prediction while preserving signed edge weights, and $\hat{V}_{A,ij}$ is used for uncertainty regularization in the connectivity loss.

The global parameters are estimated from subject-level summaries of Z_{spat}: The global parameters are estimated by attention-pooling the spatial ROI embeddings and passing the pooled representation through global MLP heads:

$$(\hat{a}, \hat{b}) = f_{\text{global}}(Z_{\text{spat}}),$$

while the ROI-wise parameter c is estimated per region:

$$\hat{c}_i = f_c(Z_{\text{spat},i}).$$

Transit time is estimated using both timing-sensitive and spatially contextualized information:

$$\hat{\tau}_i^{\text{pre}} = f_\tau^{\text{pre}}\left(Z_{\text{temp},i}, Z_{\text{phase},i}\right),$$

$$\hat{\tau}_i = \hat{\tau}_i^{\text{pre}} + \eta f_\tau^{\text{ref}}\left(Z_{\text{spat},i}\right),$$

where η controls the contribution of the spatial refinement term.

C Training Objective and Loss Terms

The model is trained with a composite objective:

$$\mathcal{L} = \mathcal{L}_A + \mathcal{L}_{\text{param}} + \lambda_{\text{CSD}}\mathcal{L}_{\text{CSD}},$$

where $\mathcal{L}_A$ supervises effective connectivity, $\mathcal{L}_{\text{param}}$ supervises non-connectivity DCM parameters, and $\mathcal{L}_{\text{CSD}}$ enforces spectral forward-model consistency.

C.1 Effective Connectivity Loss

The effective connectivity objective combines edge-wise, structural, sign, uncertainty, residual, and contrastive terms:

$$\mathcal{L}_A = \lambda_{\text{mse}}\mathcal{L}_{\text{mse}} + \lambda_{\text{corr}}\mathcal{L}_{\text{corr}} + \lambda_{\text{sign}}\mathcal{L}_{\text{sign}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{resid}}\mathcal{L}_{\text{resid}} + \lambda_{\text{cont}}\mathcal{L}_{\text{cont}}.$$

Weighted edge-wise MSE. The weighted MSE emphasizes strong reference edges and incorporates confidence from the reference posterior variance V_A:

$$\mathcal{L}_{\text{mse}} = \frac{\sum_{b,i,j} w_{b,ij} \left(\hat{A}_{b,ij} - A_{b,ij}\right)^2}{\sum_{b,i,j} w_{b,ij}}.$$

The weight is

$$w_{b,ij} = [1 + (w_{\text{strong}} - 1)\mathbf{1}\{|A_{b,ij}| > \tau_A\}]\, c_{b,ij},$$

where τ_A defines strong edges and $c_{b,ij}$ is the confidence weight derived from the inverse reference variance.

Correlation loss. To preserve subject-wise global connectivity structure, we use a Pearson correlation loss between vectorized estimated and reference matrices:

$$\mathcal{L}_{\text{corr}} = \frac{1}{B}\sum_{b=1}^{B}\left[1 - \rho\left(\text{vec}(\hat{A}_b), \text{vec}(A_b)\right)\right],$$

where $\text{vec}(\cdot)$ denotes vectorization of all matrix entries.

Sign-consistency loss. To encourage correct excitatory/inhibitory polarity on strong edges, we apply sign supervision only where $|A_{b,ij}| > \tau_A$:

$$\mathcal{L}_{\text{sign}} = \frac{1}{|\Omega|}\sum_{(b,i,j)\in\Omega}\left[\tanh(\alpha\hat{A}_{b,ij}) - \text{sign}(A_{b,ij})\right]^2,$$

where

$$\Omega = \{(b,i,j) : |A_{b,ij}| > \tau_A\}.$$

The factor α controls the sharpness of the smooth sign approximation.

Uncertainty/KL loss. The edge head predicts both a mean $\hat{A}_{ij}$ and variance $\hat{V}_{ij}$, defining

$$q(A_{ij}) = \mathcal{N}(\hat{A}_{ij}, \hat{V}_{ij}).$$

Given a reference distribution

$$p(A_{ij}) = \mathcal{N}(A_{ij}, V_{A,ij}),$$

we regularize the estimated distribution using the element-wise Gaussian divergence $\text{KL}(q \,\|\, p)$:

$$\mathcal{L}_{\text{KL}} = \frac{1}{BR^2}\sum_{b,i,j}\frac{1}{2}\left[\log\frac{V_{A,b,ij}}{\hat{V}_{b,ij}} + \frac{\hat{V}_{b,ij} + (\hat{A}_{b,ij} - A_{b,ij})^2}{V_{A,b,ij}} - 1\right].$$

This regularizes the estimated edge uncertainty relative to the reference posterior variance.

Subject-specific residual loss. The residual loss is not a BOLD reconstruction loss. Instead, it encourages the model to recover subject-specific deviations from the batch-level connectivity mean. We define centered matrices:

$$A_b^c = A_b - \frac{1}{B}\sum_{b'=1}^{B} A_{b'}, \qquad \hat{A}_b^c = \hat{A}_b - \frac{1}{B}\sum_{b'=1}^{B}\hat{A}_{b'}.$$

The residual loss is

$$\mathcal{L}_{\text{resid}} = \frac{\left\|\text{vec}(\hat{A}^c) - \text{vec}(A^c)\right\|_2^2}{\|\text{vec}(A^c)\|_2^2 + \epsilon}.$$

This term penalizes failure to capture inter-subject variability in effective connectivity.

Connectivity contrastive loss. In addition to edge-wise and structural supervision, we align estimated and reference connectivity matrices at the subject level. Let

$$\bar{a}_b = \mathrm{vec}(\hat{A}_b), \qquad a_b = \mathrm{vec}(A_b).$$

We normalize these vectors and define cosine similarity

$$\mathrm{sim}(\bar{a}_b, a_{b'}) = \frac{\bar{a}_b^\top a_{b'}}{\|\bar{a}_b\|_2 \|a_{b'}\|_2}.$$

The contrastive loss is

$$\mathcal{L}_{\mathrm{cont}} = -\frac{1}{B}\sum_{b=1}^{B} \log \frac{\exp(\mathrm{sim}(\bar{a}_b, a_b)/\tau_{\mathrm{cont}})}{\sum_{b'=1}^{B} \exp(\mathrm{sim}(\bar{a}_b, a_{b'})/\tau_{\mathrm{cont}})},$$

where τ_{cont} is the contrastive temperature. This encourages each estimated connectivity matrix to be closer to its matched reference matrix than to reference matrices from other subjects in the batch.

C.2 Non-Connectivity Parameter Losses

We supervise non-connectivity DCM parameters with mean-squared error losses. For the global spectral parameters,

$$\mathcal{L}_a = \|\hat{a} - a\|_2^2, \qquad \mathcal{L}_b = \|\hat{b} - b\|_2^2.$$

For the ROI-wise parameters c and transit time τ ,

$$\mathcal{L}_c = \frac{1}{R}\sum_{i=1}^{R} (\hat{c}_i - c_i)^2, \qquad \mathcal{L}_\tau = \frac{1}{R}\sum_{i=1}^{R} (\hat{\tau}_i - \tau_i)^2.$$

The aggregate non-connectivity parameter loss is

$$\mathcal{L}_{\mathrm{param}} = \lambda_a \mathcal{L}_a + \lambda_b \mathcal{L}_b + \lambda_c \mathcal{L}_c + \lambda_\tau \mathcal{L}_\tau.$$

C.3 Analytic Cross-Spectral Density Loss

We define the analytic spectral forward operator

$$\mathcal{S}_{\mathrm{DCM}} : \theta \mapsto \mathcal{G}_{\mathrm{DCM}}(\omega; \theta),$$

which maps DCM parameters to the model-implied BOLD cross-spectral density. This operator is the frequency-domain counterpart of the DCM forward model described in Appendix A. In particular, the nonlinear hemodynamic state equations are linearized around the resting flow to obtain a frequency response from latent neuronal activity to BOLD signal fluctuations.

Under linear neuronal dynamics with endogenous drive $u(t)$, outlined in Eq. (1), the frequency-domain transfer function is

$$T_A(\omega) = (i\omega I - A)^{-1}.$$

Let $\mathcal{G}_u(\omega)$ denote the cross-spectral density of the endogenous neuronal drive. The latent neuronal CSD is then

$$\mathcal{G}_x(\omega; \theta) = T_A(\omega)\, \mathcal{G}_u(\omega)\, T_A(\omega)^H,$$

where $(\cdot)^H$ denotes conjugate transpose.

The linearized hemodynamic model contributes a frequency response

$$H_{\mathrm{hemo}}(\omega; \theta),$$

which maps neuronal fluctuations to BOLD fluctuations. The resulting model-implied BOLD CSD is

$$\mathcal{G}_{\mathrm{DCM}}(\omega; \theta) = H_{\mathrm{hemo}}(\omega; \theta)\, \mathcal{G}_x(\omega; \theta)\, H_{\mathrm{hemo}}(\omega; \theta)^H + \mathcal{G}_e(\omega),$$

where $\mathcal{G}_e(\omega)$ denotes the observation-noise spectrum. In our notation,

$$\hat{\mathcal{G}} = \mathcal{S}_{\mathrm{DCM}}(\hat{\theta}) = \mathcal{G}_{\mathrm{DCM}}(\omega; \hat{\theta}).$$

The spectral consistency loss compares the model-implied and reference spectra in log-amplitude space:

$$\mathcal{L}_{\mathrm{CSD}} = \left\| \log(|\mathcal{S}_{\mathrm{DCM}}(\hat{\theta})| + \epsilon) - \log(|\mathcal{G}_{\mathrm{ref}}| + \epsilon) \right\|_2^2,$$

where ϵ is a small constant for numerical stability and $\mathcal{G}_{\mathrm{ref}}$ denotes the reference cross-spectral density (for training, we set $\mathcal{G}_{\mathrm{ref}} = \mathcal{S}_{\mathrm{DCM}}(\theta)$ since the ground-truth θ is known; for fine-tuning on real datasets, one could in principle set $\mathcal{G}_{\mathrm{ref}}$ as the cross-spectral density of the empirical data). This loss term encourages inferred parameters to reproduce the spectral structure of the observed BOLD dynamics, rather than only matching parameter targets independently.

Algorithm 1 BASIL training step

Input: Mini-batch $Y \in \mathbb{R}^{B\times R\times T}$, targets $\theta = \{A, a, b, c, \tau\}$, reference variance V_A, reference spectrum $\mathcal{G}_{\text{ref}}$
Output: Predictions $\hat{\theta} = \{\hat{A}, \hat{a}, \hat{b}, \hat{c}, \hat{\tau}\}$

Preprocessing

Standardize each ROI time series:

$$\tilde{Y}_{b,i,:} \leftarrow \frac{Y_{b,i,:} - \mu_{b,i}}{\sigma_{b,i} + \varepsilon}$$

Compute Hilbert-phase representation:

$$\phi_{b,i,:} \leftarrow \angle\big(\tilde{Y}_{b,i,:} + j\,\mathcal{H}(\tilde{Y}_{b,i,:})\big)$$

Representation learning

Encode temporal dynamics with Mamba:

$$Z_{\text{temp}} \leftarrow f_{\text{temp}}(\tilde{Y})$$

Encode phase-sensitive timing features:

$$Z_{\text{phase}} \leftarrow f_{\text{phase}}(\phi)$$

Contextualize ROI embeddings with the spatial Transformer:

$$Z_{\text{spat}} \leftarrow f_{\text{spatial}}(Z_{\text{temp}})$$

Compute subject-level context:

$$g_{\text{subj}} \leftarrow f_{\text{subj}}\left(\frac{1}{R}\sum_{i=1}^{R} Z_{\text{spat},i}\right)$$

Produce connectivity-conditioned ROI embeddings:

$$Z_{\text{conn}} \leftarrow f_{\text{FiLM}}(Z_{\text{spat}}, g_{\text{subj}})$$

Parameter estimation

Estimate effective connectivity:

$$A^{\text{raw}}_{ij} = f_A(Z_{\text{conn},i}, Z_{\text{conn},j}, g_{\text{subj}}), \qquad \hat{A}_{ij} = s\,\tanh(A^{\text{raw}}_{ij})$$

Estimate edge uncertainty:

$$\ell_{ij} = f_V\left(Z_{\text{conn},i}, Z_{\text{conn},j}, g_{\text{subj}}\right), \qquad \hat{V}_{A,ij} = \text{softplus}(\ell_{ij}) + \epsilon_V$$

Estimate global and ROI-wise non-connectivity parameters:

$$(\hat{a}, \hat{b}) \leftarrow f_{\text{global}}(Z_{\text{spat}}), \qquad \hat{c}_i \leftarrow f_c(Z_{\text{spat},i})$$

Estimate transit time:

$$\hat{\tau}^{\text{pre}}_i \leftarrow f^{\text{pre}}_\tau(Z_{\text{temp},i}, Z_{\text{phase},i}), \qquad \hat{\tau}_i \leftarrow \hat{\tau}^{\text{pre}}_i + \eta\, f^{\text{ref}}_\tau(Z_{\text{spat},i})$$

Forward-model consistency

Compute model-implied spectrum:

$$\hat{\mathcal{G}} \leftarrow \mathcal{S}_{\text{DCM}}(\hat{\theta})$$

Optimization

Compute effective-connectivity loss:

$$\mathcal{L}_A = \lambda_{\text{mse}}\mathcal{L}_{\text{mse}} + \lambda_{\text{corr}}\mathcal{L}_{\text{corr}} + \lambda_{\text{sign}}\mathcal{L}_{\text{sign}} + \lambda_{\text{KL}}\mathcal{L}_{\text{KL}} + \lambda_{\text{resid}}\mathcal{L}_{\text{resid}} + \lambda_{\text{cont}}\mathcal{L}_{\text{cont}}$$

Compute non-connectivity parameter loss:

$$\mathcal{L}_{\text{param}} = \lambda_a\mathcal{L}_a + \lambda_b\mathcal{L}_b + \lambda_c\mathcal{L}_c + \lambda_\tau\mathcal{L}_\tau$$

Compute spectral consistency loss:

$$\mathcal{L}_{\text{CSD}} = \left\|\log(|\hat{\mathcal{G}}| + \varepsilon) - \log(|\mathcal{G}_{\text{ref}}| + \varepsilon)\right\|_2^2$$

Form total objective:

$$\mathcal{L} = \mathcal{L}_A + \mathcal{L}_{\text{param}} + \lambda_{\text{CSD}}\mathcal{L}_{\text{CSD}}$$

Update f_ϕ using AdamW

return $\hat{\theta}$

Interpretation of Loss Components. Each loss term contributes a distinct inductive bias:

- $\mathcal{L}_{\text{mse}}$: edge-wise magnitude recovery, with emphasis on strong and high-confidence edges;
- $\mathcal{L}_{\text{corr}}$: preservation of subject-wise global connectivity structure;
- $\mathcal{L}_{\text{sign}}$: recovery of excitatory/inhibitory edge polarity;
- $\mathcal{L}_{\text{KL}}$: regularization of estimated edge uncertainty against reference posterior variance;
- $\mathcal{L}_{\text{resid}}$: recovery of subject-specific deviations from population-average connectivity;
- $\mathcal{L}_{\text{cont}}$: subject-level contrastive alignment between estimated and reference connectivity matrices;
- $\mathcal{L}_{\text{param}}$: supervision of non-connectivity DCM parameters;
- $\mathcal{L}_{\text{CSD}}$: spectral forward-model consistency under the analytic DCM operator.

For completeness, Algorithm 1 provides the full BASIL training step.

D Summary of Model Components and Hyperparameters

Tables S2 and S3 summarize the main architectural, loss, and optimization hyperparameters used in BASIL .

Table S2: Architecture components and latent representations.

Symbol / Name	Value	Description
Data and representations		
B	16	Training batch size.
R	100	Number of ROIs.
T	1200	Number of fMRI time points per ROI.
F	32	Number of frequency bins in the CSD.
d_{time}	96	Dimension of temporal ROI embeddings returned by the temporal encoder.
d_{Mamba}	64	Internal hidden dimension of the Mamba sequence model.
d_{node}	128	Dimension of spatial ROI embeddings in Z_{spat} and Z_{conn}.
d_{phase}	24	Dimension of Hilbert-phase embeddings Z_{phase}.
Z_{temp}	$B \times R \times 96$	Mamba-based temporal ROI embeddings.
Z_{phase}	$B \times R \times 24$	Multiscale Hilbert-phase embeddings.
Z_{spat}	$B \times R \times 128$	Spatially contextualized ROI embeddings.
Z_{conn}	$B \times R \times 128$	FiLM-conditioned ROI embeddings used for effective-connectivity estimation.
Temporal and phase encoders		
f_{temp}	Mamba	Temporal encoder for long-context ROI dynamics.
Mamba layers	4	Number of selective state-space layers.
Bidirectional temporal encoding	true	Uses both forward and backward temporal encoders.
f_{phase}	multiscale CNN	Phase encoder applied to Hilbert-phase time series.
Phase CNN kernels	$3, 9, 21$	Parallel convolutional kernels for multiscale phase features.
Phase fusion kernel	5	Convolutional kernel used to fuse multiscale phase features.
Phase temporal pooling	adaptive average	Pools the fused phase features over time to produce $Z_{\text{phase}} \in \mathbb{R}^{B\times R\times 24}$.
Spatial encoder and conditioning		
W_{proj}	$d_{\text{time}} \times d_{\text{node}}$	Learned linear projection that maps Z_{temp} from d_{time} to d_{node}.
f_{spatial}	Transformer	Spatial Transformer applied to the projected temporal ROI embeddings $\widetilde{Z}_{\text{temp}}$.
Spatial layers	4	Number of Transformer encoder layers.
Attention heads	8	Number of spatial self-attention heads.
Transformer MLP ratio	4	Feedforward dimension is $4d_{\text{node}}$.
Spatial dropout	0.1	Dropout in spatial Transformer and parameter heads.
f_{subj}	MLP	Subject-summary network applied to mean-pooled spatial ROI embeddings Z_{spat}.
f_{FiLM}	FiLM	Subject-conditions Z_{spat} to produce Z_{conn}.
FiLM modulation scale	0.3	Scales FiLM feature-wise modulation.
$f_{\text{edge}} = \{f_A, f_V\}$	directed edge MLP	Effective-connectivity head applied to ordered pairs $(Z_{\text{conn},i}, Z_{\text{conn},j})$ and g_{subj}.
f_A	edge MLP	Estimates the scalar pre-activation for the posterior mean of each edge
f_V	variance edge MLP	Estimates the variance pre-activation l_{ij}, converted to a positive variance by softplus.
Edge-head hidden size	128	Hidden dimension used in both f_A and f_V.
s / A output scale	0.05 init.	Learnable scale in $\hat{A} = \tanh(A^{\text{raw}}) \cdot s)$.
Minimum A variance	10^{-6}	Added after softplus in the edge variance head.
f_{global}	MLP	Estimates global spectral parameters $(\hat{a}, \hat{b})$.
Output dimension of $\hat{a}$	2	Two-dimensional global parameter a.
Output dimension of $\hat{b}$	2	Two-dimensional global parameter b.
f_c	ROI-wise MLP	Estimates one c_i per ROI.
f_τ^{pre}	MLP	Estimates preliminary transit time from $(Z_{\text{temp},i}, Z_{\text{phase},i})$.
Transit pre-head input size	120	$d_{\text{time}} + d_{\text{phase}} = 96 + 24$.
f_τ^{ref}	MLP	Refines transit time using $Z_{\text{spat},i}$.
Transit refinement scale η	0.25	Weight on the spatial refinement term.

E Baseline Methods

We compare our model against mechanistic, statistical, neural, and neuroscience-specific causal discovery baselines. The selected methods cover complementary assumptions: explicit biophysical inversion, predictive causality, neural temporal modeling, adversarial generative modeling, conditional-independence discovery, and neural perturbational inference.

E.1 Baseline Descriptions

Regression Dynamic Causal Modeling (rDCM). Regression Dynamic Causal Modeling (rDCM) [14] estimates effective connectivity by inverting a linearized neuronal and hemodynamic generative model of fMRI data. It provides a mechanistic Bayesian reference model that doesn't require supervised training, albeit it introduces several simplifying assumptions listed in Table 1.

Table S3: Loss and optimization hyperparameters.

Symbol / Name	Value	Description
Effective-connectivity loss		
λ_{mse}	1.0	Weight for variance-weighted MSE loss $\mathcal{L}_{\mathrm{mse}}$.
λ_{sign}	2.0	Weight for sign-consistency loss $\mathcal{L}_{\mathrm{sign}}$.
λ_{KL}	10^{-3}	Weight for edge-uncertainty KL regularization $\mathcal{L}_{\mathrm{KL}}$.
λ_{resid}	1.5	Weight for subject-specific residual/deviation loss $\mathcal{L}_{\mathrm{resid}}$.
λ_{cont}	0.5	Weight for connectivity contrastive loss $\mathcal{L}_{\mathrm{cont}}$.
τ_A	0.01	Threshold for defining strong edges in weighted MSE and sign-consistency loss.
$\lambda_{\mathrm{strong}}$	5.0	Multiplicative emphasis applied to strong reference edges.
α	20.0	Sharpness of smooth sign approximation $\tanh(\alpha \hat{A}_{ij})$.
τ_{cont}	0.1	Temperature parameter in the connectivity contrastive loss.
V_A MSE cap	10.0	Maximum normalized inverse-variance confidence weight in weighted MSE.
V_A KL cap	0.25	Upper cap applied to reference variance in the KL term.
Variance floor	10^{-6}	Lower bound used for V_A in dataset preprocessing and MSE confidence.
Non-connectivity parameter losses		
λ_a	1.0	Implicit weight for supervised MSE on global parameter a.
λ_b	1.0	Implicit weight for supervised MSE on global parameter b.
λ_c	1.0	Implicit weight for supervised MSE on ROI-wise parameter c.
λ_τ	2.5	Weight for final transit-time loss $\mathcal{L}_\tau$.
Spectral forward-model consistency		
$\mathcal{S}_{\mathrm{DCM}}$	analytic operator	Differentiable spectral forward operator mapping estimated parameters to model-implied BOLD CSD.
λ_{CSD}	0.1	Maximum weight for spectral consistency loss $\mathcal{L}_{\mathrm{CSD}}$.
CSD warmup	30 epochs	Effective CSD weight is linearly warmed up over 30 epochs.
CSD loss type	log-magnitude MSE	Loss between $\log(\lvert\hat{\mathcal{G}}\rvert + \epsilon)$ and $\log(\lvert\mathcal{G}_{\mathrm{ref}}\rvert + \epsilon)$.
CSD ϵ	10^{-6}	Numerical stabilizer in the log-magnitude CSD loss.
Optimization and training		
Optimizer	AdamW	Optimizer used to update all learnable parameters.
Learning rate	2×10^{-4}	AdamW learning rate used in the active run.
Weight decay	10^{-4}	AdamW weight decay.
Maximum epochs	100	Maximum number of training epochs.
Gradient clipping	0.5	Maximum gradient norm used for training stability.

Neural Granger Causality (NGC). Neural Granger Causality [32, 23] replaces linear autoregression with neural predictors to capture nonlinear temporal dependencies. We used the component-wise LSTM formulation, where one LSTM predictor is trained per target time series. Causal influence is extracted from the norms of the input-to-hidden weights: if the input weights from source region j to the target-specific predictor for region i are driven to zero under sparsity regularization, then j is not considered Granger-causal for i. This follows the cLSTM formulation in which the model contains one LSTM network per time series and extracts a Granger-causality matrix from the input weight norms. We used hidden size 64, maximum lag 5, group-lasso sparsity regularization, learning rate 10^{-3}, and 20–50 epochs. This baseline tests whether nonlinear predictive causality is sufficient for recovering directed interactions, but it remains a predictive rather than biophysical model.

EC-GAN. EC-GAN [33] formulates effective-connectivity estimation as an adversarial generative problem. The generator is composed of a set of effective-connectivity generators, one per brain region, which synthesize each region's fMRI time series from other regions and noise using causal parameters. A shared discriminator distinguishes real from generated fMRI signals. The learned causal parameters define the directed effective-connectivity matrix. Following the original formulation, the generator is based on a structural-equation-model-inspired design, the discriminator is a fully connected network with dropout, and sparsity is encouraged through a penalty on the causal parameter matrix. Edges are obtained by thresholding the learned causal parameters according to the maximum allowed number of parent nodes. This baseline tests whether adversarial distribution matching can recover effective connectivity from fMRI, although adversarial training can be unstable and does not explicitly model hemodynamic biophysics.

Temporal Causal Discovery Framework (TCDF). TCDF [34] is a temporal causal discovery framework based on attention-equipped convolutional neural networks. It trains one causal prediction network per target time series, where each network receives all observed time series as input and estimates a single target. Candidate causes are identified from attention scores, validated by an

intervention-based causal validation step, and assigned delays by interpreting convolutional filters. In our fMRI implementation, TCDF is preceded by an LSTM-based deconvolution module that estimates latent neural activity from BOLD signals. The architecture consists of a bidirectional LSTM with two layers and hidden size 64, followed by a TCDF module with dilated depthwise separable 1D convolutions using kernel size 4, two layers, dilation factors $[1, 4]$, and hidden dimension 64. MLP heads estimate coupling strength and delay. We trained with batch size 32, learning rate 10^{-3}, weight decay 10^{-5}, and 50 epochs. This baseline evaluates whether attention-based temporal CNNs can recover nonlinear lagged causal relationships from fMRI-derived neural activity.

CaLLTiF. CaLLTiF [26] is a whole-brain fMRI causal discovery method based on PCMCI-style conditional independence testing. It begins from a complete lagged parent set and performs momentary conditional independence tests over lagged and contemporaneous variables. The method outputs a lag-resolved graph and p-value matrix, which are then summarized across lags to obtain a binary directed adjacency matrix. Thus, unlike our model, CaLLTiF does not directly estimate signed or continuous coupling strengths; it estimates whether each directed edge is present or absent, with p-values serving as statistical evidence. We used partial-correlation conditional independence tests, $\tau_{\min} = 0$, $\tau_{\max} = 3$, significance level $\alpha = 0.01$, and a maximum conditioning-set size of 5 for computational stability. This baseline tests whether constraint-based causal discovery can recover directed fMRI interactions, but it is computationally sensitive to the number of ROIs and conditioning variables.

Neural Perturbational Inference (NPI). NPI [22] learns a neural surrogate model of brain dynamics and estimates effective connectivity by virtually perturbing each region. The surrogate predicts the next neural state from previous states; after training, one ROI is perturbed and the change in predicted activity across all target ROIs is averaged over time to define one row of the effective-connectivity matrix. We used the MLP surrogate model with input dimension $3R$, hidden dimension $2R$, latent dimension $0.8R$, and output dimension R, trained using mean squared error with Adam. Perturbation strength was set to half the standard deviation of the BOLD signal, following the original NPI setup. NPI provides a signed weighted EC matrix and is therefore closer to our target than binary causal discovery methods, but it requires training a surrogate model and does not explicitly incorporate DCM-style biophysical constraints.

These baselines collectively evaluate whether causal structure can be recovered through mechanistic modeling (rDCM), statistical predictability (GC), nonlinear temporal modeling (NGC, TCDF), generative modeling (EC-GAN), and neuroscience-specific causal discovery frameworks (CaLLTiF, NPI). In contrast, our approach integrates temporal dynamics, generative modeling, and domain-specific constraints within a unified framework, enabling scalable and physiologically grounded causal inference from fMRI time series.

E.2 Baseline Hyperparameters

We summarize the architectures and key hyperparameters of all baseline methods used in our experiments in Table S4.

E.3 Graph-level properties and sign-level recovery

We additionally examined whether each method preserves the sign of directed interactions (i.e., excitatory vs. inhibitory connections). Figure S1 reports row-normalized sign-confusion summaries for all edges with nonzero true coupling, separating true negative and true positive interactions and showing, within each true-sign class, the fraction assigned to each estimated sign.

BASIL shows the most balanced sign recovery across both excitatory and inhibitory interaction types, correctly recovering the majority of positive edges while retaining substantially higher true-negative recovery than most baselines. TCDF also recovers positive interactions well, but shows a stronger bias toward positive estimates. In contrast, NPI, EC-GAN, and rDCM exhibit larger sign asymmetries, indicating that their estimated matrices may reproduce some graph-level statistics while distorting the excitatory–inhibitory balance of the effective-connectivity matrix.

Next, beyond edge-wise recovery metrics reported in the main text, we also asked whether the estimated connectivity matrices preserve higher-order graph properties of the true effective-connectivity matrices. Note that CaLLTiF and NGC are excluded from this analysis because CaLLTiF returns a binary directed adjacency matrix, whereas NGC returns a weighted but non-signed adjacency matrix; neither therefore provides a signed, weighted effective-connectivity matrix. Figure S2 compares the distributions of four graph-level summaries across subjects: average edge weight, modularity, maximum betweenness, and spectral radius. For each method, split radial violins show the true matrix

Table S4: Baseline architectures and hyperparameters.

Method	Architecture / estimator	Main hyperparameters
rDCM	Bayesian inversion of linearized neuronal–hemodynamic DCM	TR $= 0.72$s
GC	Vector autoregressive model with statistical Granger testing	Maximum lag $p = 5$, significance $\alpha = 0.05$, pairwise or multivariate VAR depending on ROI setting.
NGC	Component-wise LSTM with sparsity regularization	One LSTM per target ROI, hidden size 64, lag/window length 5, group-lasso penalty, learning rate 10^{-3}, 20–50 epochs.
EC-GAN	GAN with SEM-inspired effective-connectivity generators and shared discriminator	One generator per ROI, fully connected discriminator with dropout, Gaussian noise input, sparsity penalty on causal matrix, parent threshold controlled by maximum parent count.
TCDF	LSTM deconvolution + attention-based dilated temporal CNN	BiLSTM hidden size 64, 2 layers; TCDF kernel size 4, 2 convolutional layers, dilations $[1, 4]$, hidden size 64, batch size 32, learning rate 10^{-3}, weight decay 10^{-5}, 50 epochs.
CaLLTiF	PCMCI/MCI conditional-independence testing with lagged parent sets	ParCorr test, $\tau_{\min} = 0$, $\tau_{\max} = 3$, $\alpha = 0.01$, max conditioning size 5. Output is a binary directed graph with p-values.
NPI	MLP surrogate brain with virtual ROI perturbations	Input $3R$, hidden $2R$, latent $0.8R$, output R, Adam optimizer, MSE loss, perturbation strength $0.5\times$ BOLD standard deviation. R is the total number of ROIs.

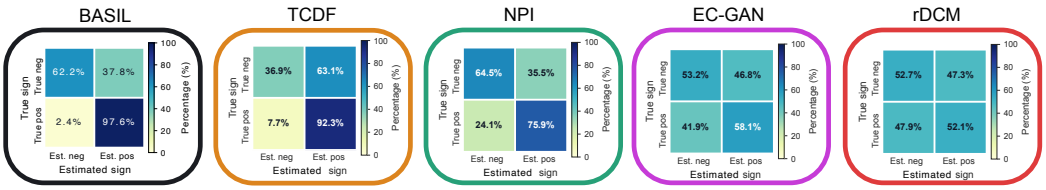


Figure S1: **Sign-confusion analysis across methods.** For each method, rows are normalized by the true edge sign, so diagonal entries correspond to the true negative rate (TNR; true negative edges estimated as negative) and true positive rate (TPR; true positive edges estimated as positive). BASIL shows the most balanced recovery of signed interactions, whereas several baselines exhibit stronger biases toward one estimated sign, particularly positive-edge assignments.

distribution in the pink half and the estimated matrix distribution in the green half, with the median and interquartile range indicated in black.

Because these metrics have different numerical ranges and units, values were normalized before visualization. Specifically, each metric was normalized using a robust global range computed from the 2nd–98th percentiles across all methods, subjects, and true/estimated matrices. Values outside this range were clipped to $[0, 1]$. Thus, the radial coordinate should be interpreted as a normalized relative scale for comparing distributional agreement, rather than as the raw metric value.

Overall, BASIL shows the closest distributional alignment between estimated and reference graph properties across the evaluated metrics followed by TCDF. This indicates that the proposed model does not merely recover individual edges, but also preserves system-level characteristics of the effective-connectivity graph, including edge-strength scale, modular organization, centrality structure, and spectral dynamics. In contrast, most baselines exhibit clear shifts in one or more graph summaries, suggesting that accurate edge-wise estimation does not necessarily guarantee preservation of the broader network topology.

To complement these cohort-level distributional comparisons, Figure. S3 presents paired subject-level comparisons between the true and BASIL estimated graph metrics. Each point represents one subject, and Pearson correlation coefficients and mean absolute errors quantify subject-specific recovery.

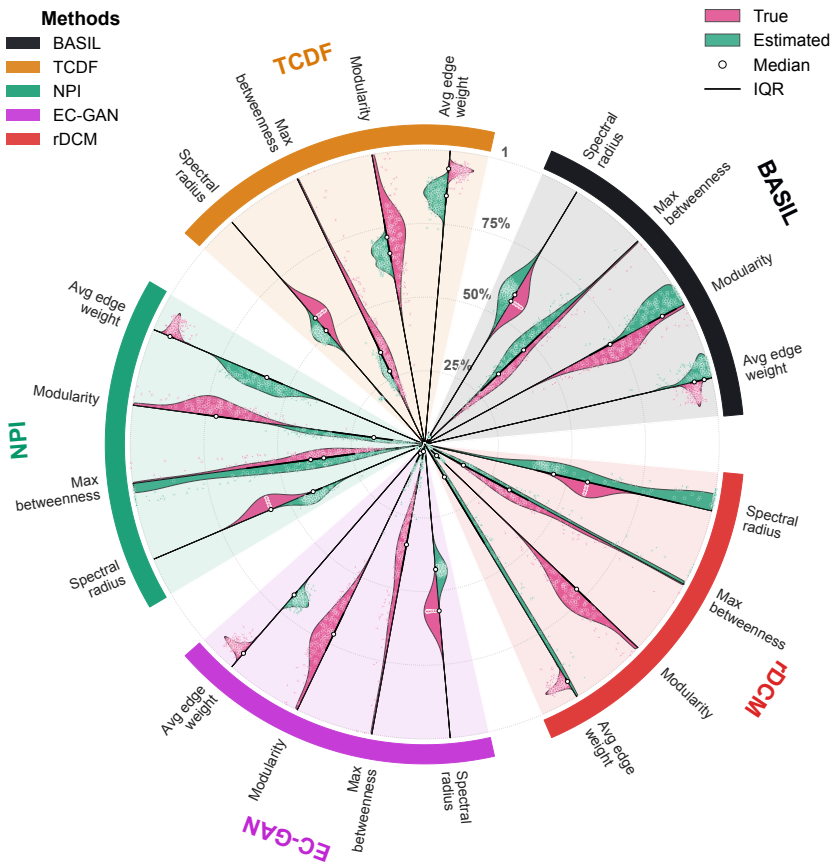


Figure S2: **Graph-level comparison of true and estimated effective-connectivity matrices.** Split radial violins compare subject-wise graph metrics computed from the true and estimated connectivity matrices for each method. Pink halves show reference matrices and green halves show estimated matrices; white markers and black bars indicate medians and interquartile ranges. Metrics were robustly normalized to $[0, 1]$ using the 2nd–98th percentile range across all methods and true/estimated matrices, enabling comparison of distributional agreement across graph summaries with different scales.

Finally, we evaluated whether each method preserves subject-to-subject similarity structure in the inferred connectivity matrices. Figure S4A compares the mean inter-subject correlation of directed effective-connectivity edges across methods. BASIL most closely matches the inter-subject similarity of the true matrices, indicating that it preserves not only average connectivity structure but also variability across individuals. We further compute this similarity within each canonical functional network block (Fig. S4B,C). Radar plots show the average pairwise subject correlation computed within each functional network block (VIS, SOM, DAN, VAN, LIM, FPN, DMN), comparing true (pink) and estimated (green) matrices. BASIL closely follows the true network-wise profile across VIS, SOM, DAN, VAN, LIM, FPN, and DMN, whereas several baselines show either compressed or shifted profiles, suggesting weaker preservation of network-specific inter-subject variability.

F Ablation Study

We perform a series of ablations to evaluate the contribution of each major component of our model and to compare alternative temporal and spatial modeling choices.

Full model (BASIL , A5). Our full model integrates a Mamba-based temporal encoder for long-range dynamics, a spatial Transformer for flexible inter-regional interaction modeling, and a cross-spectral density (CSD) loss to enforce frequency-domain consistency.

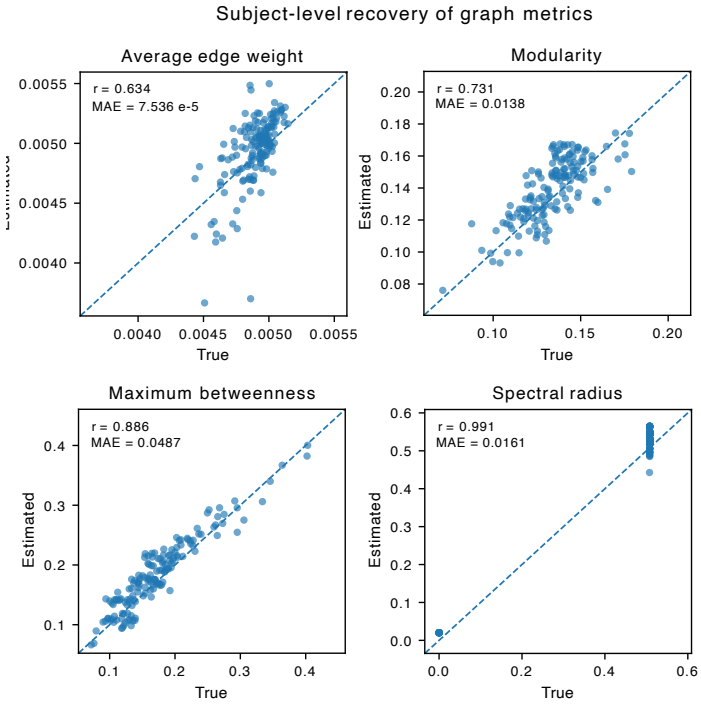


Figure S3: Subject-level recovery of graph metrics using BASIL . Scatter plots compare the true and estimated average edge weight, modularity, maximum betweenness, and spectral radius for matched subjects. The dashed diagonal indicates perfect agreement. Pearson correlation coefficients (r) and mean absolute errors (MAE) are reported for each metric.

GRU temporal encoder (A1). We replace the Mamba encoder with a bidirectional GRU while keeping the spatial Transformer unchanged. This tests whether structured state-space modeling is necessary for capturing long-range temporal dependencies. The GRU variant shows reduced performance, indicating limited ability to disentangle delayed hemodynamic effects from directed interactions.

Signed GCN spatial encoder (A2). We replace the spatial Transformer with a signed directed graph convolutional network (GCN) that models weighted, asymmetric excitatory and inhibitory interactions. The GCN uses a fully connected learnable directed adjacency matrix, initialized from small random logits and optimized jointly with the model. We decompose this adjacency into positive and negative components, which are row-normalized separately and used for signed message passing, with positive messages added and negative messages subtracted. Although this introduces biologically motivated signed propagation, the learned adjacency is shared across subjects and input-independent at inference time. Performance degrades compared to the Transformer, suggesting that a shared propagation structure is less effective than adaptive attention for capturing complex inter-regional dependencies.

No spatial encoder (A3). We remove the spatial module entirely and estimate connectivity solely from ROI-wise temporal representations. This evaluates whether inter-regional coupling can be inferred without explicit interaction modeling. The resulting performance drop confirms that spatial interaction modeling is essential.

No CSD loss (A4). We remove the cross-spectral density loss while keeping the architecture unchanged. This tests the role of frequency-domain supervision. The model exhibits weaker recovery of dynamical and spectral properties, highlighting the importance of spectral consistency in constraining biophysical parameter estimation.

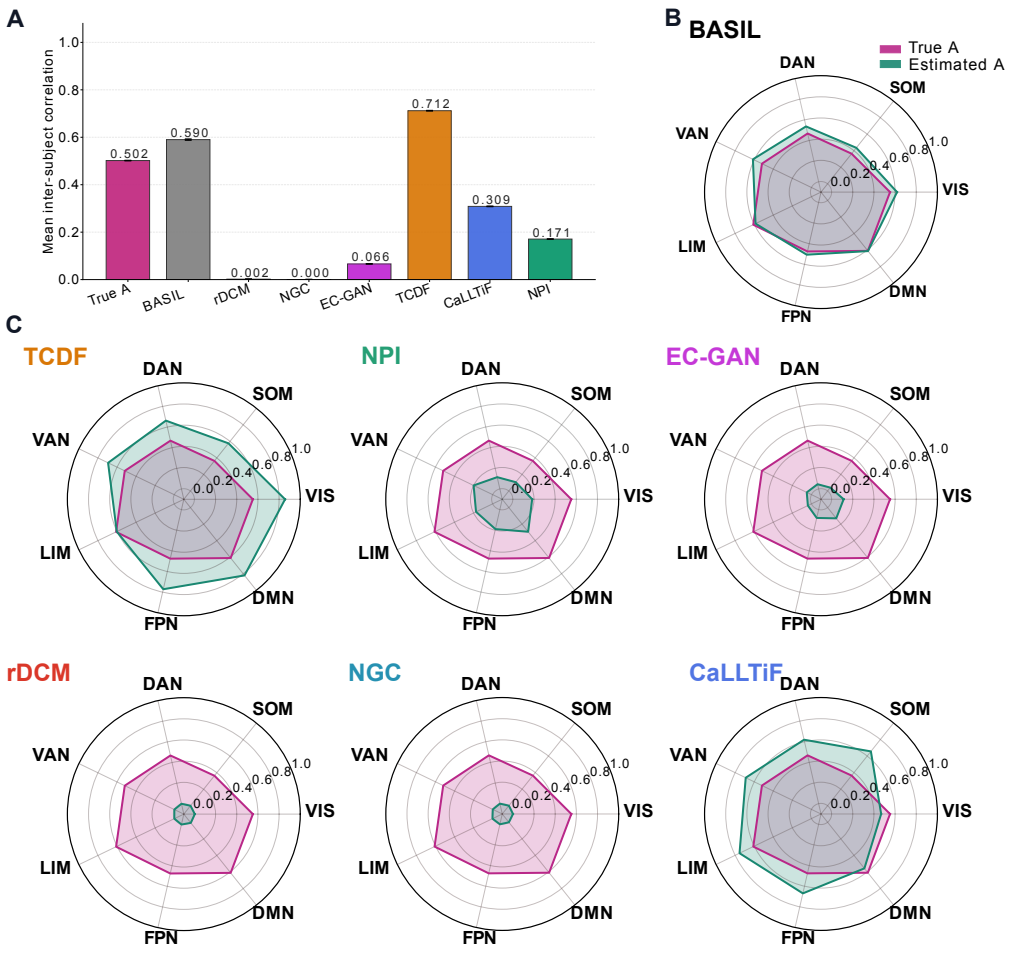


Figure S4: **Inter-subject similarity across methods and functional networks.** **(A)** Mean inter-subject correlation of directed effective-connectivity edges across methods, comparing ground truth (True A) and model estimates; black error bars denote SEM. **(B)** For BASIL , radar plots show the average pairwise subject correlation computed within each functional network block (VIS, SOM, DAN, VAN, LIM, FPN, DMN), comparing true (pink) and estimated (green) matrices. **(C)** Same analysis as (B) for baseline methods, showing how faithfully each model preserves inter-subject similarity within functional networks.

G DMN subset analysis within the Schaefer-100 benchmark

In addition to the main-text results on all 100-region Schaefer ROIs, we report a focused analysis on the *Default Mode Network (DMN)*. The DMN is a widely studied resting-state system with well-characterized mesoscale organization, making it a natural subset for detailed qualitative and network-level diagnostics. This subsection therefore serves as a compact case study that complements the full-network evaluation by providing clearer visualization of directed coupling patterns and node-level structure within a canonical network.

Setup. We restrict the analysis to ROIs assigned to the DMN under the same atlas and preprocessing pipeline used in the main text, and evaluate BASIL on the corresponding validation/test subjects with identical metrics. All reported quantities are computed on the DMN-induced subgraph, using the same conventions for directed coupling (A_{ij} corresponds to $j \to i$) and the same frequency grid for spectral diagnostics where applicable.

Parameter recovery, connectivity structure, and node-level organization. Fig. S5 provides supplementary scatter-based diagnostics and sign-consistency summaries for the DMN subset. Fig. S6 shows that BASIL preserves coarse structure within the DMN. The group-averaged estimated coupling

Table S5: Ablation study summary.

Variant	Modified / Removed Component
A1: GRU encoder	Mamba temporal modeling
A2: Signed GCN spatial encoder	Transformer-based spatial modeling
A3: No spatial encoder	Inter-ROI interaction modeling
A4: No CSD loss	Frequency-domain supervision
A5: Full BASIL	—

matrix closely matches the reference DMN matrix in terms of dominant within-network motifs and sign structure (Fig. S6A,B). Edge-wise correlation across subjects highlights which directed interactions are consistently recovered (Fig. S6C). At the node level, estimated incoming/outgoing strength profiles track the reference profiles (Fig. S6D,E), and hub-score recovery indicates that influential DMN parcels are preserved under the inferred directed graph (Fig. S6F).

These plots complement the 100-region main text results by showing that recovery trends hold in a compact network where edge patterns and ROI-wise variation can be inspected more directly. We next move beyond this simulated benchmark and evaluate BASIL on empirical HCP BOLD in a separate 24-region DMN setting.

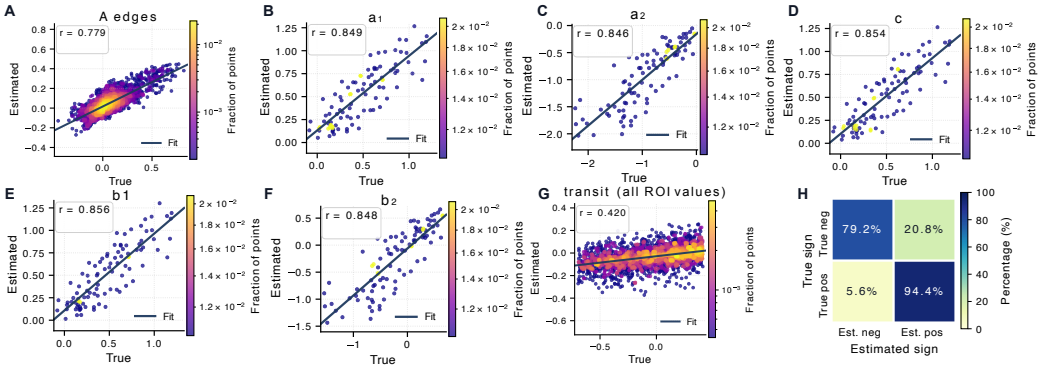


Figure S5: **DMN parameter recovery diagnostics.** DMN-only diagnostics, including scatter plots comparing estimated vs. reference quantities and sign-consistency summaries on strong directed edges.

H Empirical BOLD validation and cross-dataset adaptation

The preceding DMN analysis evaluates a subgraph of the HCP-informed simulated Schaefer-100 benchmark. Here, we provide complementary validation beyond that setting. We first evaluate a dedicated 24-region DMN model on empirical HCP BOLD, where spectral DCM remains computationally feasible, and then assess adaptation to the independent PsiConnect cohort.

Agreement with spectral DCM on real BOLD. The 24-region BASIL model was evaluated on held-out empirical HCP recordings and compared with spectral DCM estimates obtained from the same data. Agreement was moderate to strong across effective connectivity and the associated biophysical parameters:

Table S6: BASIL–spectral-DCM agreement on real HCP BOLD in the 24-region DMN.

Quantity	Pearson r
A	0.704
a_1	0.627
a_2	0.392
b_1	0.582
b_2	0.826
c	0.847
τ	0.573

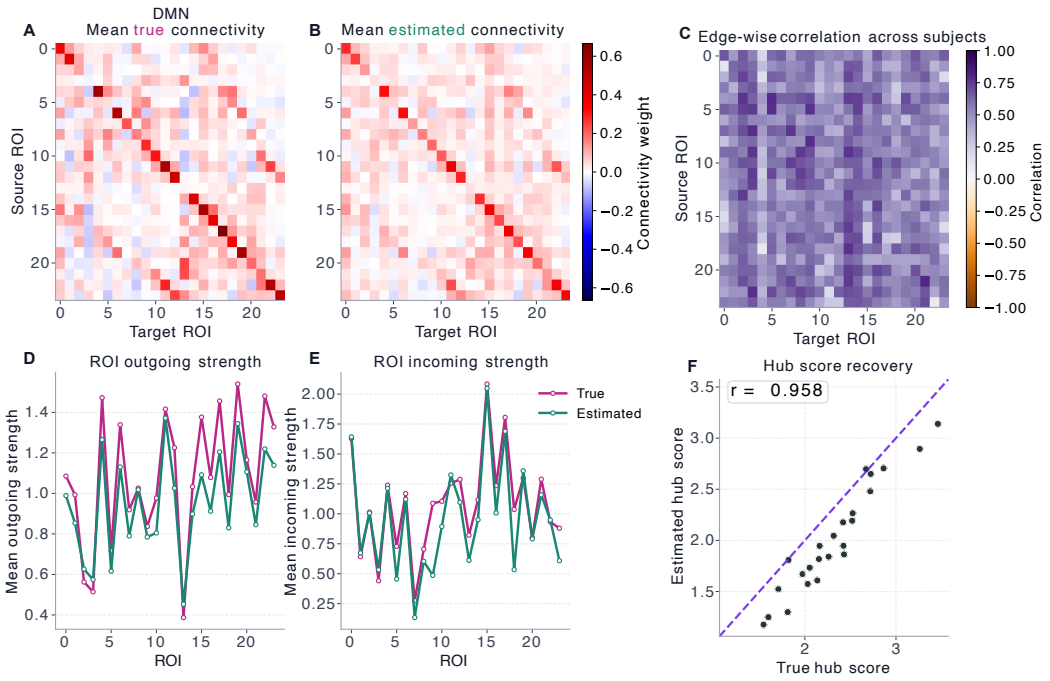


Figure S6: **DMN connectivity structure and node-level recovery. (A)** Group-averaged reference effective connectivity (A) within DMN ROIs. **(B)** Group-averaged estimated connectivity ($\hat{A}$) from BASIL , shown with the same ordering as (A). **(C)** Edge-wise correlation across subjects for each directed connection within the DMN subgraph. **(D–E)** ROI-level outgoing and incoming strength profiles averaged across all validation subjects (true vs. estimated). **(F)** Hub-score recovery comparing true and estimated hub scores within DMN.

Split-half reliability and cross-half generative consistency. To test whether estimates from empirical BOLD were reproducible across independent data segments, each held-out subject's recording was divided into two non-overlapping halves and BASIL was applied separately to each half. We quantified reproducibility using across-subject Pearson correlation, ICC$(3,1)$, and mean absolute error (MAE), together with within-subject spatial correlation for regional or edge-wise quantities.

Table S7: Split-half reliability on real HCP BOLD in the 24-region DMN.

Parameter	**Across-subject** r	**Median ICC**$(3,1)$	**MAE**	**Within-subject** r
A, off-diagonal	0.909	0.909	0.0040	0.912 ± 0.003
a_1	0.778	0.769	0.0729	–
a_2	0.342	0.342	0.0442	–
b_1	0.648	0.646	0.1183	–
b_2	0.765	0.760	0.1834	–
c	0.675	0.666	0.1094	0.968 ± 0.014
τ	0.449	0.459	0.0752	0.843 ± 0.083

Effective connectivity showed particularly strong reproducibility, including 99.4% sign agreement. Reliability was parameter-specific: a_1 and b_2 were also comparatively stable, whereas a_2 and τ showed weaker across-subject reproducibility. These results assess practical stability rather than establishing structural identifiability or uniqueness of the inverse solution.

We further used the complete parameter set inferred from one half to predict the empirical cross-spectral density (CSD) of the other:

Matched cross-half predictions were substantially stronger than shuffled subject pairings (0.905 ± 0.055 versus 0.648 ± 0.005; permutation $p = 0.0099$), indicating subject-specific spectral consistency in independent empirical BOLD segments.

Table S8: Cross-half CSD prediction from parameters inferred on real BOLD.

Cross-half prediction	CSD correlation
Half 1 → Half 2	0.901 ± 0.059
Half 2 → Half 1	0.908 ± 0.050
Bidirectional mean	0.905 ± 0.055

Downstream utility of effective connectivity. To test whether BASIL's directed matrices retain useful subject-level information beyond agreement with spectral DCM, we predicted gender using only the BASIL-estimated A matrices from real HCP BOLD in 200 held-out subjects. BASIL effective connectivity, rDCM effective connectivity, and conventional functional connectivity were evaluated using identical subject-level folds, classifier architecture, training procedure, and metrics over five repeats of stratified five-fold cross-validation.

Table S9: Matched gender-classification comparison in 200 held-out HCP subjects.

Connectivity representation	Accuracy	Balanced accuracy	ROC-AUC
BASIL effective connectivity	0.833	0.844	0.896
rDCM effective connectivity	0.628	0.634	0.694
Functional connectivity	0.733	0.747	0.747

On this matched proof-of-principle task, BASIL-derived effective connectivity outperformed both rDCM effective connectivity and undirected functional connectivity, providing complementary evidence that the inferred directed matrices retain downstream subject-level information.

Cross-dataset adaptation on PsiConnect. Finally, we assessed whether the HCP-pretrained representation could be adapted to the independent PsiConnect resting-state fMRI cohort ($N = 62$). Following the HCP procedure, simulations were generated from a PsiConnect-informed parameter distribution. HCP-pretrained fine-tuning and training from random initialisation were evaluated using identical repeated five-fold subject-level cross-validation, repeated three times, with all simulations from each empirical subject retained within the same partition and final metrics computed from pooled out-of-fold predictions.

Table S10: Parameter recovery on PsiConnect-informed simulations.

Parameter	HCP-pretrained → PsiConnect	PsiConnect from scratch
A, off-diagonal	0.598	0.643
a_1	0.573	0.715
a_2	0.534	0.790
b_1	0.604	0.785
b_2	0.591	0.634
c	0.567	0.738
τ	0.402	0.510

Fine-tuning retained meaningful recovery across the parameter set without retraining the complete model, although dataset-specific training from scratch was generally stronger. Because both experiments retain the same DCM simulation family, this analysis supports adaptation across independently informed parameter distributions rather than simulator-independent validation.

I Computational scalability

A central limitation of classical DCM inversion is computational scaling with network size. Here, we compare runtime only against biophysical DCM variants, since these are the methods that estimate DCM-style neuronal/hemodynamic parameters in addition to effective connectivity. In particular, stochastic DCM (sDCM) [43] relies on iterative variational inference and repeated evaluation of frequency-domain quantities that involve matrix inversions and updates over all regions, causing runtime to grow rapidly as the number of ROIs increases. This makes sDCM impractical for modern whole-brain parcellations, where $N \geq 100$ is common, and therefore excludes it from our main-text baseline comparisons.

Fig. S7 reports wall-clock runtime as a function of ROI count for three representative regimes (small: $N \in \{3, 10, 20\}$; moderate-to-large: $N = 100$). As expected, sDCM exhibits steep growth and becomes infeasible at $N = 100$ (marked as not completed), whereas the amortized BASIL

performs inference in a single forward pass and remains tractable at $N = 100$. For completeness, we also report rDCM, which is designed to improve scalability relative to classical DCM by forgoing the estimation of hemodynamic parameters and introducing several simplifying assumptions listed in Table 1. rDCM is faster than BASIL at $N = 100$ in this runtime comparison, but BASIL remains tractable while also providing biophysical parameter estimation.

These results underscore the practical advantage of amortized inference for cohort-scale analyses: once trained, BASIL avoids iterative per-subject optimization and enables cortex-wide effective connectivity estimation at ROI counts where classical sDCM inversion is prohibitively slow. We implement BASIL in PyTorch and train on a single NVIDIA H100 80GB GPU.

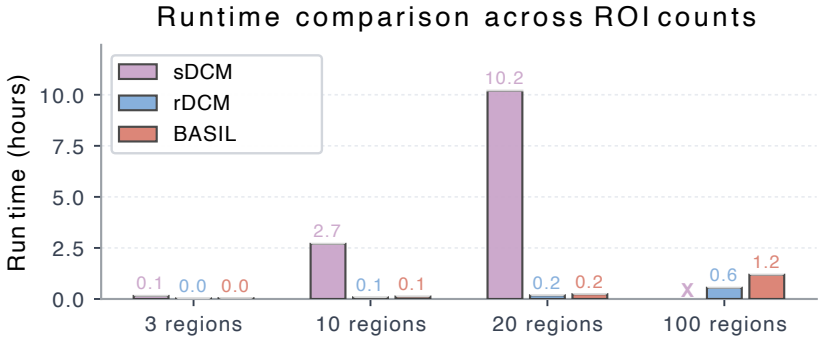


Figure S7: **Runtime scaling with ROI count.** Wall-clock runtime (hours) for sDCM, rDCM, and BASIL across increasing numbers of ROIs. sDCM becomes infeasible at $N = 100$ (not completed; marked with $\times$), whereas both rDCM and BASIL remain tractable. rDCM is faster in this comparison, while BASIL provides single-pass amortized inference within the proposed physics-informed neural framework.

J Additional Details

Seed regions were selected in the left-hemisphere core regions of each of the seven canonical resting-state networks in the locations specified as in Table S11 similar to [22].

Table S11: MNI coordinates of seed regions in seven resting-state networks.

Network	Seed region	MNI coordinate
VIS	L V2	$(-12.72, -81.95, 1.04)$
SOM	L 3b	$(-41.40, -21.76, 50.97)$
VAN	L AIP	$(-37.76, -44.82, 42.39)$
DAN	L FOP3	$(-37.44, 2.97, 12.33)$
LIM	L TGv	$(-41.18, -2.96, -44.84)$
FPN	L a9-46v	$(-40.90, 49.90, 8.64)$
DMN	L 31pv	$(-9.41, -45.69, 33.04)$

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes],

   Justification: The abstract and introduction explicitly state the paper's scope and main contributions: a physics-informed amortized inference framework for estimating directed effective connectivity and DCM-style neuronal/hemodynamic parameters from resting-state fMRI in a single forward pass. These claims are directly supported by the experiments, which report edge-level recovery, sparse-edge overlap, sign recovery, parameter recovery, graph/network-level analyses, ablations, and runtime scalability results. The paper also states key assumptions and limitations, including that supervision is derived from rDCM estimates rather than invasive ground truth, and that weakly identifiable parameters such as transit time remain challenging. Therefore, the claims are aligned with the reported results and appropriately bounded in scope.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes],

   Justification: The paper explicitly discusses limitations in the Conclusion, including the reliance on rDCM-derived supervision rather than invasive ground truth and the weaker identifiability of parameters such as transit time from resting-state BOLD alone. It also outlines future work to address these limitations through task-evoked data, multi-run/session training, and multimodal recordings.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

   Justification: [N/A]

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: The paper provides the dataset, preprocessing setup, model architecture, losses, metrics, baselines, hyperparameters, and ablation settings needed to reproduce the main results, while Appendices B-D provide further details on the method architecture, components, and hyperparameters (see Table S2-S3). We have also released an anonymous GitHub repository with the implementation and experiment code.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The experiments use the HCP dataset, which is available to researchers subject to the appropriate data-use license. We also provide an anonymous GitHub repository with the code and instructions needed to reproduce the main experiments, together with implementation and hyperparameter details in Appendices B-F.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies the HCP dataset, train/validation/test split, input length, model components, optimizer-related training details, losses, evaluation metrics, baselines, and ablation settings. Additional architecture and hyperparameter details are provided in Appendix B-D and Tables S2-S3.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The paper reports subject-level distributions and variability summaries, including per-subject points and error bars in Fig. 3B,D, subject-wise distributions in Fig. 4C, and supplementary distributional analyses in Figs. S2 and S1. These show variability across subjects and methods rather than only aggregate means.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the hardware setup (single NVIDIA H100 80GB GPU) and provide runtime scaling analyses across ROI sizes in Appendix I, enabling estimation of compute requirements.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers, CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and the work uses licensed/open-access neuroimaging data with appropriate privacy safeguards. The paper does not involve new human-subject data collection, clinical deployment, or identifiable participant information.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: In the Conclusion section, the paper discusses positive impacts, including scalable effective-connectivity analysis for neuroscience, and potential risks, including misleading interpretation from biased data or incorrect estimates. It also states that the method should be treated as a research tool rather than a clinical decision system without further validation.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A].

Justification: [N/A].

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We cite and credit the external dataset and baseline methods used in the paper, including HCP and the relevant method papers. HCP data are used under the required data-use terms, and the released code repository will specify code licensing and usage instructions.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.

- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We release a new anonymized code repository documenting the implementation, training setup, evaluation scripts, and usage instructions. The repository includes license information needed to reproduce the reported experiments.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: [N/A]

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: [N/A]

    Guidelines:

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: [N/A]