FAST NEURAL MASS MODEL INVERSION VIA AMORTIZED INFERENCE

ABSTRACT

Understanding how brain regions influence one another through effective connectivity (EC) is important for explaining brain function. Dynamic Causal Modeling (DCM) estimates these interactions using biologically grounded generative models, but requires iterative model inversion for each observation. We investigate whether this inversion can be amortized by training neural networks on simulated brain dynamics and then reusing the learned inverse model for new observations. We evaluate two amortized estimators: an EC temporal convolutional network (EC-TCN) for constrained connectivity recovery, and SparseGraphInverse, a shared edge-scoring model designed to scale to larger networks. We test these models using (1) whole-brain Jansen–Rit simulations, (2) an independent benchmark generated using the Statistical Parametric Mapping version 12 (SPM) ERP DCM forward model, and (3) auditory EEG. We further examine how recovery changes as the number of potential connections increases. In particular, we distinguish between identifying the target regions whose incoming coupling changes between conditions and identifying which source–target connections carry these changes. On the 200-case SPM benchmark, EC-TCN reaches an area under the precision-recall curve (AUPRC) of 0.593 ± 0.005 (five training seeds), 37% above DCM (0.433; 95% CI 0.402–0.465), with lower magnitude error (0.316 versus 0.484) and better-calibrated 95% intervals (0.93 versus 0.36 coverage). Inference takes 2.71ms per case (batch of 200) versus 79.13s for DCM, roughly $29{,}000\times$ faster. Amortized inversion can therefore provide fast, EC recovery when the set of possible connections is constrained.

1 INTRODUCTION

Effective connectivity (EC) describes directed interactions among neuronal populations. Dynamic causal modeling (DCM) estimates these interactions by inverting a neural mass model (NMM) that links latent neuronal dynamics to observed neurophysiological signals (Friston et al., 2003; David & Friston, 2003; Kiebel et al., 2008). For electroencephalography (EEG) and magnetoencephalography (MEG), neural mass models (NMMs) provide a mechanistic description of how population-level interactions generate measured responses, making inferred connectivity interpretable within *a priori* specified region of interest (ROI) of the brain (David & Friston, 2003; Kiebel et al., 2008). DCM and NMM approaches have been used extensively for electrophysiological connectivity analysis (Friston et al., 2003; Ursino et al., 2007) and network inference (Ditlevsen et al., 2023). Unlike statistical measures such as Granger causality and vector autoregressive models, which characterize whether past activity in one signal predicts another (Granger, 1969; Barnett & Seth, 2014; Shojaie & Fox, 2022), DCM estimates coupling parameters within generative model of neuronal dynamics.

In practice, DCM for EEG and MEG is generally applied to a relatively small number of ROIs. Variational inversion jointly estimates the model parameters, but its computational cost increases with model dimensionality as the number of regions and connectivity parameters grows, limiting the practical scalability of DCM to large networks (Friston et al., 2003; Kiebel et al., 2008).

The harder question is whether the underlying connectivity can be recovered reliably. Different connectivity changes between regions can generate very similar EEG signals, making the inverse problem ambiguous, and a model trained on simulations may also fail when new observations differ from those encountered during training (Tolley et al., 2024). Thus, scaling connectivity inference is not only a computational problem: we must determine when connectivity can be recovered, what causes recovery to fail, and whether reliable inference remains possible beyond predefined networks.

Amortized inversion of neural generative models is already established: once trained, the same estimator can be reused for new observations (Cranmer et al., 2020; Lueckmann et al., 2021). Re-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

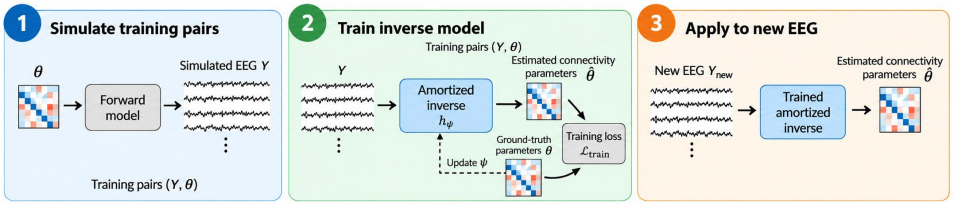


Figure 1: Overview of amortized effective-connectivity inversion. **1)** Known connectivity parameters θ are passed through a neural mass forward model to generate simulated EEG Y, forming training pairs (Y, θ). **2)** The amortized inverse receives EEG Y as input and predicts connectivity parameters $\hat{\theta}$; the known parameters θ provide supervision through the training loss used to update the inverse model. **3)** Once trained, the amortized inverse maps new EEG Y_{new} directly to estimated connectivity parameters $\hat{\theta}$.

cent work has applied this strategy to neural and whole-brain generative models (Hashemi et al., 2023; Tolley et al., 2024; Hashemi et al., 2024), including local parameter recovery in the Jansen–Rit neural mass model (Tilwani & O'Reilly, 2025). What remains unresolved is how far reliable effective-connectivity localization can scale as the candidate connectivity space expands. Within this amortized-inference framework, we use two learned inverse models: an EC temporal convolutional network (EC-TCN) for recovery over a constrained set of connections, and SparseGraphInverse for whole-brain connections. Models are trained on simulated observation–parameter pairs and applied to new observations to estimate connectivity (see Figure 1). Our experiments address four questions that determine when amortized inference can provide a scalable alternative to repeated DCM inversion: (1) under what conditions it can reliably recover connectivity changes; (2) how recovery changes as the number of ROIs increases; (3) how amortized inference compares with DCM in accuracy, uncertainty, and computational cost on simulated data with known connectivity; and (4) what connectivity the simulation-trained estimator identifies on real auditory EEG, and how well it agrees with fitted DCM.

2 Method

2.1 Problem formulation and inference objective

We infer condition-dependent changes in EC from NMM-simulated source EEG. Candidate effects are extrinsic, modifying directed coupling between distinct regions, or intrinsic, modifying a region's local self-gain. For each candidate effect e, the ground truth is (z_e, b_e), where $z_e \in \{0, 1\}$ indicates its presence and $b_e \in \mathbb{R}$ is its signed modulation, with $b_e = 0$ when $z_e = 0$. The inverse model estimates $(\hat{p}_e, \hat{b}_e)$, where $\hat{p}_e$ is the probability that the effect is present and $\hat{b}_e$ is its signed magnitude. A positive or negative b_e indicates an increase or decrease relative to baseline coupling for extrinsic effects, or baseline self-gain for intrinsic effects. Each observation also has a mechanism-family label $f \in \{1, \ldots, 6\}$: null (no modulation), forward only, backward only, bidirectional (forward and backward without intrinsic modulation), intrinsic only, or combined (forward, backward, and intrinsic modulation). The EC-TCN family classifier and the SPM DCM model space use this same definition of the combined family in the real-EEG analysis. Conventional DCM uses iterative variational inversion for each new time series. We instead train an amortized inverse model h_ψ to estimate all candidate effects directly from a new observation. This incurs an upfront simulation cost T_{sim} and training cost T_{train}, followed by inference at $T_{\text{net}} \ll T_{\text{VL}}$, where T_{VL} is DCM's per-observation inversion time. Computational speed does not resolve identifiability: parameters that produce indistinguishable observations cannot be distinguished by either method.

2.2 NMM as an EEG Generative Model

Condition-specific EEG responses, $\mathbf{Y}_c$, are generated using 68 coupled standard Jansen–Rit (JR) NMMs, one for each of the 68 regions of the Desikan–Killiany brain parcellation (Desikan et al., 2006), coupled through the ENIGMA Toolbox Human Connectome Project (HCP) normative structural connectome. Here $c \in \{\text{std}, \text{dev}\}$ denotes the standard and deviant stimulus conditions. We define the condition-specific generative model as:

$$\dot{\mathbf{x}}_c(t) = \mathbf{F}\big(\mathbf{x}_c, \mathbf{u}_c, \boldsymbol{\theta}_c(t), \mathbf{A}^{(c)}(t)\big) + \boldsymbol{\omega}_c(t), \quad \mathbf{Y}_c = \mathbf{G}\big(\mathbf{x}_c, \boldsymbol{\theta}_c(t)\big) + \boldsymbol{\varepsilon}_c, \tag{1}$$

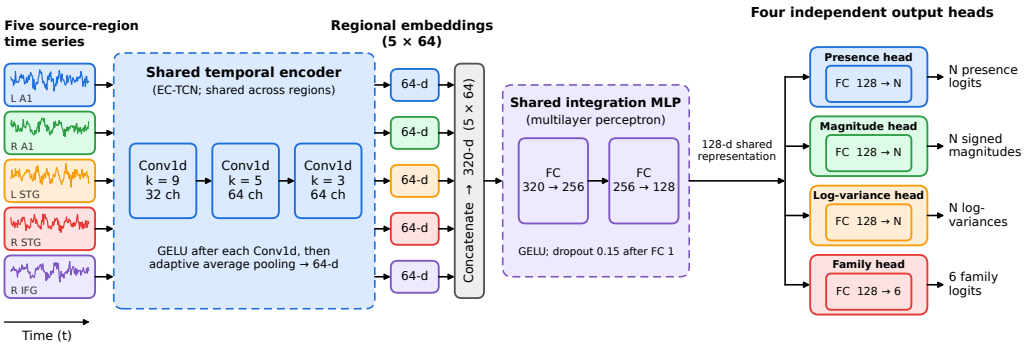


Figure 2: **EC-TCN architecture.** Five source-region time series are encoded with shared temporal convolutions, integrated into a common representation, and mapped to independent prediction heads for each edge candidate-effect inference.

where $\mathbf{F}$ specifies the neuronal dynamics, $\mathbf{G}$ the observation (source-space EEG), $\mathbf{x}_c$ the system states, $\mathbf{u}_c$ the external inputs, $\boldsymbol{\theta}_c(t)$ the local parameters to be inferred, $\mathbf{A}^{(c)}(t)$ the inter-regional coupling matrix, and $\boldsymbol{\omega}_c$ and $\boldsymbol{\varepsilon}_c$ the process and measurement noise. Here, and in the rest of this article, a *case* is equivalent to an EEG recording, but can be generated through simulation or recorded from an actual participant.

Condition-dependent effects are represented as a shared baseline plus a deviant-specific modulation that begins at stimulus onset. Let $t = 0$ denote stimulus onset and let $g(t) = \mathbb{1}[t \geq 0]$. Let $\mathbf{B}$ denote the case-specific connectivity modulation matrix, formed from whichever of the six extrinsic connectivity candidates are active for that case (zero at every other region pair). The coupling matrix for $c \in \{\mathrm{std}, \mathrm{dev}\}$ is

$$\mathbf{A}^{(c)}(t) = \mathbf{A}^{(0)} + \delta_{c,\mathrm{dev}}\, g(t)\, \mathbf{B}, \tag{2}$$

where $\mathbf{A}^{(0)}$ is the baseline inter-regional coupling matrix (the row-normalized structural connectome) and $\delta_{c,\mathrm{dev}}$ equals 1 for the deviant condition and 0 for the standard condition. Thus, both conditions use $\mathbf{A}^{(0)}$ during the burn-in and pre-stimulus interval, and the deviant condition uses $\mathbf{A}^{(0)} + \mathbf{B}$ from stimulus onset onward. Intrinsic effects are implemented analogously, as a gated perturbation of regional self-gain within $\boldsymbol{\theta}_c(t)$. Let $\mathbf{d}$ denote the case-specific self-gain perturbation vector, formed from whichever of the five intrinsic candidates are active for that case (zero at every other region). Then

$$\boldsymbol{\theta}_c(t) = \boldsymbol{\theta}^{(0)} + \delta_{c,\mathrm{dev}}\, g(t)\, \mathbf{d}, \tag{3}$$

where $\boldsymbol{\theta}^{(0)}$ is the baseline (condition-independent) parameter vector.

In our analyses, parameter inference targets the set $\mathcal{E}$ of $N = |\mathcal{E}| = 11$ candidate effects embedded in whole-brain dynamics. These effects involve five source regions: left and right primary auditory cortex (L_A1, R_A1), left and right superior temporal gyrus (L_STG, R_STG), and right inferior frontal gyrus (R_IFG). The candidate set comprises six extrinsic connectivity effects (three forward connections along the A1→STG→IFG hierarchy and three corresponding backward connections) together with five intrinsic regional self-gain effects. All N candidates are scored at validation and test time. Appendix A.1 provides the full differential equations and model constants. For the scalability analysis, we extend the generative model to all 68 regions, using the same onset-gated modulation and yielding $N = 1{,}394$ candidate directed connections from 4,556 possible region pairs; full construction details are provided in Appendix A.1.1.

2.3 Amortized inverse model: architecture, training, and calibration

2.3.1 Network architecture and output structure

EC-TCN: EC-TCN applies the same shared-weight temporal encoder to each of the five source-region time series (Fig. 2), using three Conv1d layers with kernel widths $9/5/3$ and $32/64/64$ channels. Adaptive average pooling produces a 64-dimensional embedding per region; concatenating the five embeddings yields a 320-dimensional representation. A shared integration multilayer perceptron (MLP), $320 \to 256 \to 128$, produces a 128-dimensional representation that feeds independent

heads for candidate-effect presence, signed magnitude, uncertainty, and mechanism family. The log-variance output is clamped to $[-7, 2]$ to prevent numerical extremes. A second EC-TCN variant is trained for the independent SPM DCM benchmark (Section 2.4.2), where both methods receive identical SPM-generated sensor-space observations rather than JR-simulated source-space signals; this variant operates on 130-dimensional input (65 sensors $\times$ two conditions, folded channel-wise) and omits the mechanism-family head, since the SPM benchmark's ground truth does not include family labels. Full architecture details for this variant are provided in Appendix A.3.

SparseGraphInverse. In parallel, we train **SparseGraphInverse**, which avoids an N-dependent output layer, where N denotes the number of candidate directed edges ($N = 1{,}394$ in the whole-brain experiment). For each region r, the shared encoder produces a dense representation $\mathbf{h}_r \in \mathbb{R}^{32}$. For each candidate edge $e = (s, t)$, with source region s and target region t, we construct an edge representation by concatenating the source and target embeddings with their elementwise product: $\mathbf{h}_e = [\mathbf{h}_s \,\|\, \mathbf{h}_t \,\|\, \mathbf{h}_s \odot \mathbf{h}_t] \in \mathbb{R}^{96}$. The same edge-level MLP is applied to every candidate edge: $(\eta_e, \hat{b}_e) = \mathrm{MLP}_\theta(\mathbf{h}_e)$, $\hat{p}_e = \sigma(\eta_e)$, $\mathrm{MLP}_\theta = \mathrm{Linear}(96, 64) \to \mathrm{GELU} \to \mathrm{Linear}(64, 2)$, where $\hat{p}_e$ estimates the probability that the true presence indicator $z_e = 1$, and $\hat{b}_e$ estimates the corresponding signed modulation b_e. Because the same MLP parameters are shared across all candidate edges, the decoder parameter count is independent of N, although predictions are produced for all N edges.

2.3.2 Training objective

Recall from §2.1 that each case $n = 1, \ldots, M$ carries a mechanism-family label $f_n \in \{1, \ldots, 6\}$ (null, forward, backward, bidirectional, intrinsic, or combined – the coarse pattern of which candidate effects are jointly active), and that each candidate effect $e = 1, \ldots, N$ (with $N = |\mathcal{E}|$) carries a ground-truth presence indicator $a_{ne} \in \{0, 1\}$ and signed change b_{ne}. Three losses, equally weighted ($\lambda_{\mathrm{pres}} = \lambda_{\mathrm{fam}} = \lambda_{\mathrm{eff}} = 1.0$), combine detection, family classification, and magnitude estimation:

$$\mathcal{L} = \underbrace{\mathrm{CE}(\hat{f}, f_n)}_{\text{family}} + \underbrace{\mathrm{BCE}(\eta_{ne}, a_{ne})}_{\text{presence}} + \frac{1}{MN} \sum_{n,e} w_{ne} \underbrace{\left[(b_{ne} - \mu_{ne})^2 e^{-\ell_{ne}} + \ell_{ne}\right]}_{\text{heteroscedastic effect regression}}, \tag{4}$$

where $\hat{f}$ is the predicted family-logit vector, η_{ne} the predicted presence logit (as in §2.1), CE is cross-entropy, BCE is binary cross-entropy on logits, and the third term is a heteroscedastic Gaussian negative log-likelihood for the effect magnitude b_{ne}, with predicted mean μ_{ne} and predicted log-variance $\ell_{ne} = \log \sigma^2_{ne}$. The family and presence labels are not redundant: f_n is a coarse, case-level summary of which mechanism pattern is active, while a_{ne} is the fine-grained, per-candidate detection target the family label is derived from; the network is supervised at both granularities. The regression term is reweighted by $w_{ne} = 0.2 + 0.8\, a_{ne}$, a fixed 5:1 emphasis on active edges chosen empirically to counteract the severe class imbalance from only 1–10 active candidates per case (not derived from an external scheme). Training uses AdamW (Loshchilov & Hutter, 2019) (lr $= 2 \times 10^{-3}$, weight decay 2×10^{-4}), batch size 64, and 100 epochs (JR) or 80 epochs (SPM). The model checkpoint with the minimum validation loss was retained for evaluation. During training, gradient norm was clipped at 3.0 to stabilize optimization. All training uses fixed random seeds per run (seeds 9411–9415 in this study).

2.3.3 Uncertainty calibration and predictive scale

After training, two validation-only calibration steps construct calibrated predictive intervals:

(i) Temperature scaling: Fit a scalar $\tau \in [0.05, 10]$ via Limited-memory Broyden–Fletcher–Goldfarb–Shanno (LBFGS) to minimize validation loss after dividing logits by τ before softmax/sigmoid. This rescales the confidence of family and presence predictions.

(ii) Per-edge variance calibration: For each edge, compute validation residuals $r_{ne} = |b_{ne} - \hat{\mu}_{ne}| / \exp(0.5\hat{\ell}_{ne})$, then set $s_e = \frac{P_{0.95}(r_{:,e})}{1.96}$, clamped to $[0.05, 20]$, where $P_{0.95}$ denotes the 95th percentile. The final predictive scale is $\hat{\sigma}_{ne} = \exp(\hat{\ell}_{ne}/2) \cdot s_e$. This combines the network's learned per-observation log-variance with a fixed per-edge recalibration factor, ensuring that $\hat{\mu}_{ne} \pm 1.96\hat{\sigma}_{ne}$ achieves approximately 95% empirical coverage on validation data. Both calibration factors are frozen before evaluation; no real data enters either step.

Table 1: Controlled EC-TCN test cohorts (500 simulations each).

Test cohort	Change from training distribution	Purpose
Matched	None	In-distribution recovery.
Measurement noise	Noise fraction $0.02 \rightarrow 0.10$ of the signal scale.	Noisier observations.
Independent process noise	Independent, rather than matched, process noise for standard and deviant.	Latent noise no longer cancels in the contrast.
Connectome mismatch	10% of structural weights multiplied by $1 + \mathcal{N}(0, 0.2^2)$, truncated at zero.	Structural misspecification.
Off-range magnitude	$\lvert b_e \rvert \sim U(0.32, 0.52)$ instead of $U(0.08, 0.30)$.	Extrapolation beyond training magnitudes.
Combined mismatch	All four shifts together.	Compound distribution shift.

2.4 EXPERIMENTS

2.4.1 CONSTRAINED EXPERIMENT: JANSEN–RIT COHORT

All 68 regions generate background dynamics, while inference is restricted to a five-region auditory network. Each of the 11 candidate effects is independently active with probability 0.38 and signed magnitude $|b_e| \sim U(0.08, 0.30)$; 12% of cases are forced to be fully null, and a case with no active effect receives one at random. The mechanism-family label (§2.3.2) is derived from the active effects, so classes are imbalanced (training: null 11.5%, forward 1.8%, backward 1.3%, bidirectional 5.0%, intrinsic 4.5%, combined 75.9%). Modulation ranges are chosen for numerical stability rather than biological interpretation, and JR parameters are sampled independently per case (Appendix A.1). We simulate standard and deviant conditions with Heun integration (1 ms step, 100 ms burn-in), with the deviant modulation beginning at stimulus onset. The input is the deviant-minus-standard difference, a 5×150 matrix at 250 Hz (-100 to 500 ms), normalized by its RMS, transformed as $\operatorname{arcsinh}(\Delta / \mathrm{RMS}(\Delta))$ to preserve sign, and standardized with training-set statistics. The dataset has 3,500 training, 500 validation, and six 500-case test cohorts spanning in-distribution and mismatch settings (Table 1).

2.4.2 BASELINE COMPARISON: SYNTHETIC EEG DCM BENCHMARK

To avoid comparing the amortized estimator on its own generative model, we construct an independent benchmark using SPM's event-related potential (ERP) DCM forward model, using the sensor-space EC-TCN variant introduced above. Both EC-TCN and DCM receive the same 65-sensor, 76-sample observations, with known ground truth (no real participants). The dataset comprises 600 training (EC-TCN only), 120 validation (EC-TCN calibration only), and 200 test cases. Active magnitudes follow $|b_e| \sim \mathrm{U}(0.15, 0.70)$, distinct from the JR generator's training range ($\mathrm{U}(0.08, 0.30)$) to prevent optimistic overlap with EC-TCN's training distribution. For each test case, DCM independently fits a complete 11-parameter model using Variational Laplace (VL); edge presence is scored by sign-symmetric posterior confidence $P_p = 1 - \Phi(0; |E_p|, C_p)$, thresholded at 0.95. For comparison, nine additional baseline methods (linear, ridge, random forest, MLP, gated recurrent unit (GRU), Transformer, VAR/Granger, and two SBI variants (Cranmer et al., 2020; Lueckmann et al., 2021; Hashemi et al., 2023)) are trained on identical splits (full configurations in Appendix A.6).

2.4.3 SCALING EXPERIMENT

To test how candidate-space size affects *edge localization*, i.e, identifying the specific directed source–target connection whose coupling changes, we expand the candidate pool (i.e., potential connections) from 11 to $N \in \{25, 50, 100, 250, 500, 1{,}394\}$ edges. Each simulated case contains 1–10 active effects (i.e., actual connections) sampled uniformly from the candidate pool, while the remaining simulation protocol is held fixed. Increasing N therefore enlarges the set of competing connectivity hypotheses from which the inverse model must identify the active edges. Per-region temporal encoders produce source embeddings, and each candidate edge is represented from its source and target embeddings and scored by a single shared edge-scoring MLP. Because the same scoring MLP is reused across candidates, its learned output parameters do not grow with N, unlike EC-TCN's candidate-specific output heads. A widened EC-TCN control is described in Appendix A.7.1. Comparing EC-TCN with SparseGraphInverse therefore tests whether declining edge-localization performance can be attributed to output-capacity scaling alone. We separately test *source competition* by holding the total candidate count fixed while varying how many candidate source regions can connect to each target. Simply *can we identify the active edge $e = (s, t)$, or only the affected target region t?* We evaluate exact edge recovery and, conditional on the true target connection set, source identification; these metrics are defined over inferred connectivity changes

rather than changes in the target region's observed activity. Full scaling results and controls are in Appendix A.7.

2.4.4 REAL AUDITORY EEG: FROZEN TRANSFER WITHOUT RETRAINING

We apply two simulation-trained EC-TCN models, without retraining, to the passive auditory oddball task in OpenNeuro ds006480 (68 participants) (Kim et al., 2026). The first is the edge-level model trained on the 11-effect cohort above. The second is a family-level model trained on a separate cohort of 2,000 simulations in which the six mechanism classes are sampled uniformly (ten seeds, probabilities averaged). We use the control condition, which has no threat of unpredictable shock, and the input is the deviant-minus-standard response in that condition (the "passive-control contrast"). Sixty-four participants have complete recordings in all four conditions and are preprocessed and source-reconstructed as described in Appendix A.8; DCM was fit to all 68. We estimate source activity in the five auditory regions with MNE's dSPM (Dale et al., 2000; Gramfort et al., 2013) (default parameters) on the fsaverage template head (MNE's fsaverage head-to-MRI transform, ico-4 source space, three-layer BEM), using an ad hoc noise covariance because participant MRIs are unavailable. Because the models are trained on simulations, we first check each recording against the simulation distribution. Each recording is summarized by a 17-dimensional descriptor of temporal, spectral, amplitude, and inter-regional correlation statistics (median and IQR across regions; Appendix A.8.1). It is flagged as out-of-domain if its Mahalanobis distance from a model's training simulations exceeds the 99th percentile of that model's validation simulations. The gate admits 63 of 64 recordings for the edge-level model and 42 of 64 for the family-level model. On the 63 admitted recordings, we compare EC-TCN and SPM DCM signed effects for the 11 candidate edges, standardized within each method because the two parameterizations are on different scales. We compare standardized EC-TCN and SPM DCM edge estimates and separately compare their mechanism-family distributions. Because edge-level ground truth is unavailable, we interpret the real-EEG results at the level of mechanism families (null, forward, backward, bidirectional, intrinsic, combined). For the family comparison, we apply SPM random-effects Bayesian model selection (BMS) (Stephan et al., 2009; Penny et al., 2010) to the subject-level DCM fits, restricted to the 42 participants admitted for the family-level model, and report expected family and exceedance probabilities. An exploratory 15-edge graph analysis is reported in Appendix A.8.2.

2.5 EVALUATION METRICS AND STATISTICAL PROCEDURES

We report AUPRC over flattened case–candidate pairs and Recall@true-k, the fraction of true active effects among the top k predictions. For magnitude and uncertainty of active effects, we report MAE, sign accuracy, and 95% coverage: $\mathrm{MAE} = \mathbb{E}|\hat{b}_{ne} - b_{ne}|$, $\mathrm{SignAcc} = \Pr[\mathrm{sgn}(\hat{b}_{ne}) = \mathrm{sgn}(b_{ne})]$, and $|b_{ne} - \hat{b}_{ne}| \leq 1.96\hat{\sigma}_{ne}$. Coverage measures empirical predictive calibration and is not equivalent to DCM posterior uncertainty.

2.5.1 COMPARATIVE STATISTICS AND REPORTING

Simulated benchmarks: Results are reported as mean $\pm$ SD across independent training seeds (five for JR and SPM; three for scaling). **EC-TCN vs. DCM:** Metric differences and 95% CIs are estimated from 10,000 bootstrap resamples of the 200 test cases, preserving all 11 edges within each case. **Real EEG:** Family-level agreement is reported descriptively because ground truth is unavailable. **Runtime:** Frozen EC-TCN is evaluated on a single CPU thread at batch size 200 using 50 timed runs after 10 warm-ups, excluding model loading. DCM timing includes model construction, dipole fitting, and variational inversion; thus, the ratio compares end-to-end inference cost rather than identical operations.

3 RESULTS

3.1 CONSTRAINED EXPERIMENT: JANSEN–RIT COHORT

Across five independently trained EC-TCN models, edge recovery was strong on the matched cohort and remained stable under independent process noise and connectome mismatch (Table 2). Increased measurement noise was the one perturbation that degraded both ranking and calibration

(AUPRC .905 → .864; 95% coverage .939 → .871).Off-range magnitudes and the combined shift had a smaller effect on edge ranking and sign recovery but substantially degraded magnitude estimation and empirical coverage. Thus, the principal failure under extrapolation was magnitude calibration rather than edge detection. Recovery also depended on mechanism: intrinsic effects were most recoverable, followed by forward and backward effects. Generator-level effect-signature analysis showed that forward and backward effects produced responses of similar magnitude relative to null variability. However, different connections terminating at the same region could produce highly similar observation signatures. For example, $R_A1 \rightarrow R_STG$ and $R_IFG \rightarrow R_STG$ produced nearly identical mean signatures despite originating from different regions (Appendix A.2, Table 4). This suggests that poorer edge localization can arise from difficulty distinguishing among alternative incoming connections, rather than from uniformly weaker backward effects.

Table 2: Constrained JR inversion under matched and shifted test conditions. Mean ± SD across five independently trained EC-TCN models; 500 simulations per cohort. AUPRC and F1 pool the 11 candidate effects; sign accuracy, MAE, and coverage are evaluated on ground-truth active effects.

Cohort	AUPRC	F1	Sign acc.	MAE	95% cov.
Matched	$.905 \pm .006$	$.780 \pm .009$	$.899 \pm .004$	$.077 \pm .001$	$.939 \pm .004$
Measurement noise	$.864 \pm .014$	$.730 \pm .018$	$.893 \pm .011$	$.086 \pm .004$	$.871 \pm .021$
Independent process noise	$.898 \pm .006$	$.779 \pm .010$	$.904 \pm .008$	$.078 \pm .002$	$.939 \pm .004$
Connectome mismatch	$.902 \pm .008$	$.772 \pm .009$	$.907 \pm .007$	$.075 \pm .002$	$.946 \pm .004$
Off-range magnitude	$.888 \pm .006$	$.772 \pm .004$	$.912 \pm .007$	$.255 \pm .008$	$.265 \pm .010$
Combined shift	$.889 \pm .008$	$.768 \pm .007$	$.908 \pm .008$	$.267 \pm .008$	$.232 \pm .014$

3.2 BASELINE COMPARISON: SYNTHETIC EEG DCM BENCHMARK

On the independent 200-case SPM benchmark, EC-TCN achieved better predictive recovery than the evaluated SPM DCM configuration across all reported metrics (Figure 3). A paired case-level bootstrap placed the AUPRC advantage at 0.169 (95% CI 0.134–0.202). The advantage was strongest for intrinsic effects; the forward- and backward-specific intervals included zero. Linear, random-forest, and neural baselines also exceed DCM in pooled AUPRC; we compare all other models in Appendix A.6.

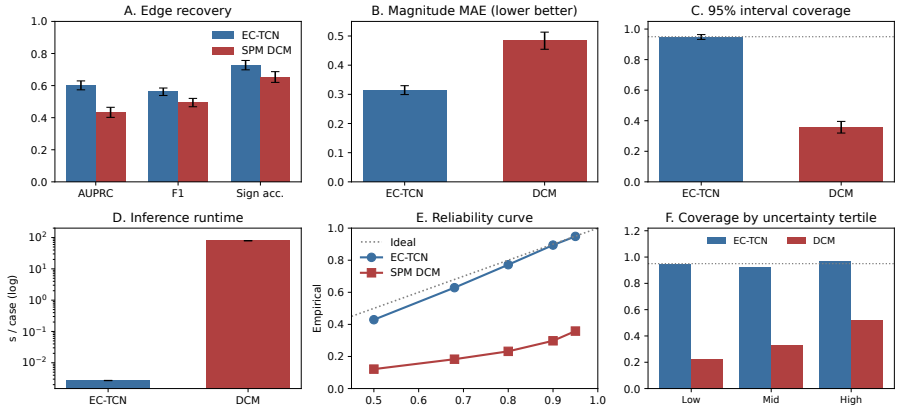


Figure 3: EC-TCN versus SPM DCM on the same 200-case synthetic SPM benchmark. Bars show the single locked EC-TCN model (AUPRC 0.601; the abstract quotes the five-seed mean, 0.593) with paired-bootstrap 95% intervals. Runtime compares frozen EC-TCN with per-case DCM inversion.

Conditional calibration showed that EC-TCN maintained approximately nominal 95% coverage across its uncertainty tertiles (0.926–0.973), whereas the evaluated DCM intervals showed substantially lower coverage (0.224–0.521); in both methods, interval width and MAE increased with reported uncertainty (Appendix A.4, Table 5). The computational difference was substantially larger: frozen EC-TCN inference required 2.71 ms per case compared with 79.13 s for DCM, approximately 2.92×10^4-fold faster per case. Including EC-TCN's one-time training and evaluation cost reduced this to a 321× end-to-end difference for the 200-case benchmark. As a convergence control, we refit 40 cases with progressively higher DCM iteration limits; all fits converged at the highest

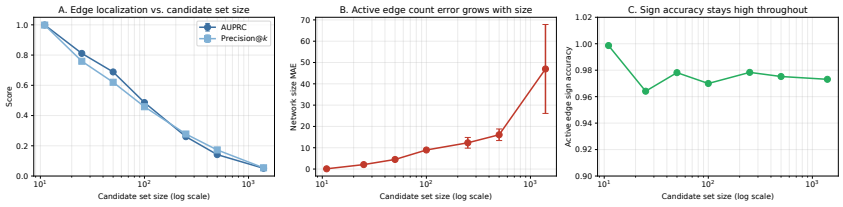


Figure 4: **Candidate-space scaling.** SparseGraphInverse performance as the candidate pool expands from $N = 11$ to 1,394, with the architecture and full 68-region observation held fixed. Curves show mean $\pm$ SD across three independently trained seeds.

limit, while DCM AUPRC changed little, indicating that insufficient iterations do not explain the observed performance difference (Appendix A.5; Table 6).

3.3 SCALING EXPERIMENT

Holding the full 68-region observation and SparseGraphInverse architecture fixed, edge localization deteriorated as the directed candidate pool expanded (Figure 4; Appendix Table 8). AUPRC and precision@k fell approximately linearly against log candidate count, while the mean absolute error in the estimated number of active edges increased. Sign accuracy on ground-truth active edges remained high, indicating that scaling primarily impaired identification of *which* edges were active, rather than estimation of their signs on the true active edges. The decline persisted under matched random candidate pools (Appendix A.7.2), showing that it is not specific to deterministic top-N selection, and additional stimulation conditions or temporal samples did not restore localization (Appendix A.7.3). Replacing the decoder gave the same result (Appendix A.7.5). A fixed-$N = 500$ source-competition control further separated target localization from source attribution: increasing the number of candidate sources per target sharply reduced edge-level recovery, while target-level AUPRC remained near 0.99. Even when the true target was provided, correct-source ranking remained close to chance (Appendix A.7.4). These results suggest that the scaling bottleneck lies in distinguishing among competing incoming pathways, not in output dimensionality, decoder capacity, or identifying the affected target.

3.4 REAL AUDITORY EEG: FROZEN TRANSFER WITHOUT RETRAINING

On the same 42 participants, EC-TCN and SPM DCM both assign probability above 0.97 to the presence of any modulation (Table 3). At the edge level, standardized EC-TCN and DCM estimates on the 63 participants were uncorrelated (participant-level $\rho = 0.03$, $p = 0.50$; group-level $\rho = -0.04$, $p = 0.92$; sign agreement 0.51), which is compatible with their different generative models and parameterizations; we therefore interpret the real-EEG results at the family level. They differ on the mechanism: SPM DCM favors the Combined family (expected probability 0.567, exceedance probability 0.998), whereas EC-TCN places most probability on Bidirectional (0.422), followed by Forward (0.311) and Combined (0.231). Both assign high probability to extrinsic modulation and forward involvement; SPM DCM's lower values (0.778 and 0.754 versus 0.984 and 0.964) mainly reflect its 0.204 weight on the intrinsic-only family, which EC-TCN almost never selects (0.002). With no ground-truth connectivity for these recordings, agreement between the methods cannot establish the accuracy of either. Recording screening and the source-reconstruction pipeline are described in Appendices A.8.1 and A.8.

4 DISCUSSION AND LIMITATIONS

On an independent SPM-generated benchmark with known ground truth, EC-TCN recovered connectivity changes more accurately and performed inference approximately 29,000$\times$ faster than the evaluated SPM DCM baseline. The scaling experiments revealed a different limitation: target localization and source attribution behaved differently as the candidate graph expanded. The affected target remained highly identifiable, whereas distinguishing among competing incoming sources became increasingly difficult. Controls over candidate selection, decoder architecture, and observation richness suggest (Appendices A.7.2–A.7.5) that this ambiguity is not explained simply by model capacity, but reflects an identifiability constraint of the NMM and experimental setting.

Table 3: Real-EEG mechanism-family comparison on the same 42 participants. EC-TCN reports mean class probability ± SD across 10 models; SPM DCM reports random-effects expected and exceedance probabilities. The two columns are different quantities, and no ground-truth connectivity exists for these recordings.

Mechanism Family	**EC-TCN Probability**	**SPM Expected**	**SPM Exceedance**
Bidirectional	0.422 ± 0.109	0.125	≈ 0
Forward	0.311 ± 0.126	0.042	≈ 0
Combined	0.231 ± 0.068	0.567	0.998
Backward	0.020 ± 0.012	0.041	≈ 0
Null	0.014 ± 0.011	0.021	≈ 0
Intrinsic	0.002 ± 0.002	0.204	0.002
Aggregate Quantities	**EC-TCN**	**SPM DCM**	
Any modulation	0.986	0.977	
Extrinsic modulation	0.984	0.778	
Forward involvement	0.964	0.754	

This highlights where amortized inference is especially useful. By shifting the main computational cost to simulation and training, it enables rapid repeated inference across many observations within a fixed, simulatable hypothesis space, making it attractive for large cohorts and high-throughput analyses. DCM remains valuable when the hypothesis space changes substantially or when an observation-specific variational posterior is required. The two methods also quantify uncertainty differently: EC-TCN is calibrated over simulated examples, whereas DCM estimates a posterior for each observation. Amortized inversion therefore provides a practical complement to conventional inference when the same generative hypothesis space is evaluated repeatedly. On real auditory EEG (Table 3), the simulation-trained model was applied without retraining to recordings that passed domain screening. EC-TCN and DCM broadly agreed on the presence of modulation and substantial extrinsic and forward involvement, but differed in their preferred mechanism family and individual edge estimates. Because ground-truth EC is unavailable for these recordings, these differences reflect the two inference approaches rather than establish that either is more accurate on real EEG.

These findings have limits. The comparison with DCM was made in controlled simulations, against a DCM baseline capped at 64 iterations; the advantage held on a subset run to convergence, but this is established for that DCM specification, not for every prior or parameterization. The results are tied to the specific JR-NMM and SPM model, their bounded parameter ranges, and the observation models evaluated here. The real-EEG evaluation uses a single auditory dataset, and EC-TCN predictive intervals and DCM posterior covariance quantify different forms of uncertainty. Within these limits, amortized inversion is effective when the candidate connectivity space is constrained and physiologically bounded. Extending precise edge localization to larger connectivity spaces will likely require additional anatomical priors or experimental constraints that help distinguish competing sources. In our experiments (Appendix A.7.3), observing all 68 regions and increasing temporal or stimulation information did not resolve this ambiguity (Appendices A.7.1 and A.7.3).

5 Conclusion

Amortized inversion offers a way to replace repeated neural-mass-model optimization with fast, direct inference while maintaining accurate effective-connectivity recovery within a constrained hypothesis space. On an independent SPM benchmark, EC-TCN showed improved connectivity recovery and substantially lower per-case inference time. The scaling experiments further separated target localization from source attribution: target-level recovery remained strong as the candidate space expanded, whereas distinguishing among competing incoming sources became more difficult. On real auditory EEG, the simulation-trained model could be applied without retraining to recordings that passed domain screening, although ground-truth connectivity is unavailable for these data. Together, these results support amortized inversion as a useful approach for repeated effective-connectivity inference and suggest that resolving ambiguity among competing sources will be important for extending the method to larger connectivity spaces.

6 EXTRAS

AI USE STATEMENT

The authors used generative AI solely for proofreading and language editing during manuscript preparation. The authors reviewed all edits and take full responsibility for the content of this publication.

ETHICS STATEMENT

Real EEG data used in this work (Section 2.4.4, Appendix A.8) come from OpenNeuro ds006480, a public, de-identified dataset released for secondary research use; no new human subjects were recruited and no identifiable data were collected, so no additional IRB review was required for this secondary analysis. All findings from this dataset are explicitly reported as exploratory: EC-TCN was trained only on simulated data, as explained in (Appendix A.8.1); we screened recordings whose statistical structure is unlike the simulations before any real data result is reported, and no clinical or diagnostic claims are made or implied. We see no other foreseeable harms, conflicts of interest, or dual use concerns arising from this work.

REPRODUCIBILITY STATEMENT

The model architectures and training procedures for EC-TCN and the exploratory whole-brain baseline (SparseGraphInverse) are specified in Section 2 and Appendices A.3 through A.9. The Jansen–Rit simulator, its constants, and the sampled parameter ranges are given in Appendix A.1. The SPM benchmark (case counts, effect magnitudes, and DCM fitting settings) is described in Section 2.4.2 and Appendices A.3 and A.5, and all baseline configurations are in Appendix A.6. The calibration procedure is in Section 2.3.3, with additional calibration results in Appendix A.4. The real-EEG pipeline (preprocessing, source reconstruction, and screening against surrogate controls) is in Appendices A.8 and A.8.1. The expanded 15-edge graph was chosen using simulation validation AUPRC only (Appendix A.8.2). Random seeds and software versions are listed in Appendix A.9. The real data are the public OpenNeuro dataset ds006480 and the Human Connectome Project connectome distributed with the ENIGMA Toolbox; rerunning the DCM baseline requires SPM and MATLAB. Runtimes were measured on an Apple-silicon CPU and depend on hardware. Code will be provided on acceptance.

REFERENCES

Lionel Barnett and Anil K Seth. The mvgc multivariate granger causality toolbox: a new approach to granger-causal inference. *Journal of neuroscience methods*, 223:50–68, 2014.

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.

Anders M Dale, Arthur K Liu, Bruce R Fischl, Randy L Buckner, John W Belliveau, Jeffrey D Lewine, and Eric Halgren. Dynamic statistical parametric mapping: combining fmri and meg for high-resolution imaging of cortical activity. *neuron*, 26(1):55–67, 2000.

Olivier David and Karl J Friston. A neural mass model for meg/eeg:: coupling and neuronal dynamics. *NeuroImage*, 20(3):1743–1755, 2003.

Rahul S Desikan, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, Bradley T Hyman, et al. An automated labeling system for subdividing the human cerebral cortex on mri scans into gyral based regions of interest. *Neuroimage*, 31(3):968–980, 2006.

Susanne Ditlevsen, Massimiliano Tamborrino, and Irene Tubikanec. Network inference in a stochastic multi-population neural mass model via approximate bayesian computation. *arXiv preprint arXiv:2306.15787*, 2023.

Karl J Friston, Lee Harrison, and Will Penny. Dynamic causal modelling. *Neuroimage*, 19(4): 1273–1302, 2003.

Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, et al. Meg and eeg data analysis with mne-python. *Frontiers in Neuroinformatics*, 7:267, 2013.

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

Meysam Hashemi, Anirudh N Vattikonda, Jayant Jha, Viktor Sip, Marmaduke M Woodman, Fabrice Bartolomei, and Viktor K Jirsa. Amortized bayesian inference on generative dynamical network models of epilepsy using deep neural density estimators. *Neural Networks*, 163:178–194, 2023.

Meysam Hashemi, Abolfazl Ziaeemehr, Marmaduke M Woodman, Jan Fousek, Spase Petkoski, and Viktor K Jirsa. Simulation-based inference on virtual brain models of disorders. *Machine Learning: Science and Technology*, 5(3):035019, 2024.

Ben H Jansen and Vincent G Rit. Electroencephalogram and visual evoked potential generation in a mathematical model of coupled cortical columns. *Biological cybernetics*, 73(4):357–366, 1995.

Stefan J Kiebel, Marta I Garrido, Rosalyn J Moran, and Karl J Friston. Dynamic causal modelling for eeg and meg. *Cognitive neurodynamics*, 2(2):121–136, 2008.

Andy Jeesu Kim, Mara Mather, Santiago Morales, and Joshua Senior. Young adult resting state and auditory oddball task eeg data, 2026. URL `https://doi.org/10.82901/nemar.on006480`.

Sara Larivière, Casey Paquola, Bo-yong Park, Jessica Royer, Yezhou Wang, Oualid Benkarim, Reinder Vos de Wael, Sofie L Valk, Sophia I Thomopoulos, Matthias Kirschner, et al. The enigma toolbox: multiscale neural contextualization of multisite neuroimaging datasets. *Nature methods*, 18(7):698–700, 2021.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of multivariate analysis*, 88(2):365–411, 2004.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

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

Will D Penny, Klaas E Stephan, Jean Daunizeau, Maria J Rosa, Karl J Friston, Thomas M Schofield, and Alex P Leff. Comparing families of dynamic causal models. *PLoS computational biology*, 6 (3):e1000709, 2010.

Paula Sanz Leon, Stuart A Knock, M Marmaduke Woodman, Lia Domide, Jochen Mersmann, Anthony R McIntosh, and Viktor Jirsa. The virtual brain: a simulator of primate brain network dynamics. *Frontiers in neuroinformatics*, 7:10, 2013.

Ali Shojaie and Emily B Fox. Granger causality: A review and recent advances. *Annual review of statistics and its application*, 9:289–319, 2022.

Klaas Enno Stephan, Will D Penny, Jean Daunizeau, Rosalyn J Moran, and Karl J Friston. Bayesian model selection for group studies. *Neuroimage*, 46(4):1004–1017, 2009.

Deepa Tilwani and Christian O'Reilly. Deep jansen-rit parameter inference for model-driven analysis of brain activity. In *Proceedings of the 7th International Conference on Advances in Signal Processing and Artificial Intelligence (ASPAI 2025)*, pp. 92–99, 2025.

Nicholas Tolley, Pedro LC Rodrigues, Alexandre Gramfort, and Stephanie R Jones. Methods and considerations for estimating parameters in biophysically detailed neural models with simulation based inference. *PLOS Computational Biology*, 20(2):e1011108, 2024.

Mauro Ursino, Melissa Zavaglia, Laura Astolfi, and Fabio Babiloni. Use of a neural mass model for the analysis of effective connectivity among cortical regions based on high resolution eeg recordings. *Biological Cybernetics*, 96(3):351–365, 2007.

A APPENDIX

The main text contains the study design, primary results, and limitations. This appendix adds simulator and model specifications, robustness checks, and supporting analyses.

A.1 JANSEN–RIT SIMULATOR DETAILS

Each of the 68 Desikan–Killiany regions is a Jansen–Rit neural mass (Jansen & Rit, 1995). Regions interact through their firing rates, weighted by structural connectivity, as in standard whole-brain models (David & Friston, 2003; Sanz Leon et al., 2013). A region contains pyramidal cells, excitatory interneurons, and inhibitory interneurons. Of its six states, x_{1i} and x_{2i} are the excitatory and inhibitory post-synaptic potentials on the pyramidal cells of region i, x_{0i} is the potential that pyramidal firing evokes in the interneurons, and x_{3i}, x_{4i}, x_{5i} are the time derivatives of x_{0i}, x_{1i}, x_{2i}. The observed source signal of region i is $x_{1i} - x_{2i}$, the net potential on its pyramidal cells, a standard proxy for the EEG a region generates. The dynamics are

$$\begin{aligned} \dot{x}_{0i} &= x_{3i}, & \dot{x}_{3i} &= H_{e,i}(t)\, a\, S(x_{1i} - x_{2i}) - 2a\, x_{3i} - a^2 x_{0i}, \\ \dot{x}_{1i} &= x_{4i}, & \dot{x}_{4i} &= H_{e,i}(t)\, a\, [p_i(t) + C_2 S(C_1 x_{0i})] - 2a\, x_{4i} - a^2 x_{1i}, \\ \dot{x}_{2i} &= x_{5i}, & \dot{x}_{5i} &= H_i\, b\, C_4 S(C_3 x_{0i}) - 2b\, x_{5i} - b^2 x_{2i}, \end{aligned}$$

where $S(v) = 2e_0/[1 + \exp\{r(v_0 - v)\}]$ converts a mean potential v into a firing rate.

The input $p_i(t)$ is a firing rate (pulses per second) arriving at the excitatory synapses of the pyramidal cells. It is not an electrical current, and it is not observed. It has four parts:

$$p_i(t) = p_0 + u_i^c(t) + \omega_i(t) + \kappa \sum_j A_{ij}^{(c)}(t)\, S(x_{1j} - x_{2j}),$$

namely the background input p_0, the stimulus-evoked input $u_i^c(t)$, process noise $\omega_i(t)$, and the firing rates of the other regions weighted by the coupling matrix $A^{(c)}$ and scaled by the global coupling κ. The process noise of Eq. (1) enters through this input, and inter-regional conduction delays are not modeled. Condition effects enter as in Eqs. (2) and (3): $A^{(c)}(t) = \widetilde{\mathbf{W}} + \delta_{c,\mathrm{dev}}\, g(t)\, \mathbf{B}$ and $H_{e,i}(t) = H_e + \delta_{c,\mathrm{dev}}\, g(t)\, d_i$, with $g(t) = \mathbb{1}[t \geq 0]$. The modulation is additive on the coupling, as in the bilinear form of DCM (Friston et al., 2003). Both conditions use baseline values during burn-in and before stimulus onset. Here $\widetilde{\mathbf{W}}$ is the 68×68 normative structural connectome of the Human Connectome Project distributed with the ENIGMA Toolbox (Larivière et al., 2021), with its diagonal set to zero and each row scaled to sum to one, so that every region receives the same total incoming weight; $\widetilde{W}_{ij}$ is the influence of region j on region i.

We use the canonical Jansen–Rit constants: excitatory and inhibitory synaptic gains $H_e = 3.25$ mV and $H_i = 22$ mV; inverse synaptic time constants $a = 100\ \mathrm{s}^{-1}$ and $b = 50\ \mathrm{s}^{-1}$; connectivity constants $C = 135$, $C_1 = C$, $C_2 = 0.8C$, $C_3 = C_4 = 0.25C$ (average numbers of synaptic contacts between the populations); and sigmoid parameters $e_0 = 2.5\ \mathrm{s}^{-1}$ (half the maximum firing rate), $v_0 = 6$ mV (potential at half-maximum rate), and $r = 0.56\ \mathrm{mV}^{-1}$ (slope).

For each simulation, five parameters are drawn: deviant gain $\gamma \sim U(0.85, 1.25)$, deviant latency $\tau \sim U(-5, 8)$ ms, global coupling $\kappa \sim U(12, 30)$, background input $p_0 \sim U(95, 145)$ pulses/s, and process-noise standard deviation $\sigma_\omega \sim U(0.05, 0.5)$ pulses/s. The stimulus input $u_i^c(t)$ is a 20-ms rectangular pulse of amplitude 50 pulses/s delivered to left and right A1. The deviant pulse is scaled by γ and shifted by τ; the standard pulse has $\gamma = 1$ and $\tau = 0$. The process noise $\omega_i(t)$ is Gaussian white noise with standard deviation σ_ω, drawn afresh at every 1-ms step and shared between the standard and deviant conditions of a pair. The background-input range is centered on the canonical value of 120 pulses/s. We checked the ranges with 496 simulations (random draws, all corners of the parameter box, and stronger-modulation test draws): all were numerically stable, with source amplitudes below 18 mV. Background input below roughly 100–106 pulses/s places a region below its oscillation threshold, so about 13% of random draws (39 of 300) show only the evoked response without spontaneous rhythm.

Measurement noise ε is added to the trial-averaged responses as Gaussian noise with standard deviation $0.02\, s/\sqrt{n_c}$, where s is the RMS amplitude of the simulated signal, 0.02 is the single-trial

noise level (2% of s), and $n_{\mathrm{std}} = 168$ and $n_{\mathrm{dev}} = 36$ are the numbers of standard and deviant trials averaged in a recording. The deviant response is therefore noisier than the standard one, as in the real data.

A.1.1 WHOLE-BRAIN EXTENSION OF THE GENERATIVE MODEL

To probe scope beyond the five-region hypothesis-driven task, we extend the same generative model (Eqs. (1)–(2)) to all $R = 68$ Desikan–Killiany regions. With 68 regions and no self-connections, each region can connect to 67 others, giving $68 \times 67 = 4{,}556$ possible directed connections. The ENIGMA connectome has $N = 1{,}394$ nonzero off-diagonal entries, which form the candidate edge pool for $\mathbf{B}$ (Eq. (2)). Unlike the five-region task, no intrinsic self-gain candidates (Eq. (3)) are included. For each simulated case, $k \sim \mathrm{Uniform}\{1, \ldots, 10\}$ candidates are drawn uniformly at random from this pool, each with a signed strength $b \sim \pm U(0.08, 0.30)$, and used to construct that case's $\mathbf{B}$; all remaining entries of $\mathbf{B}$ are zero. The stimulus is a 20-ms pulse of amplitude $U(35, 65)$ pulses/s delivered to 1–3 randomly chosen regions and is identical in both conditions; the global coupling $\kappa \sim U(12, 30)$, background input $p_0 \sim U(95, 145)$ pulses/s, and process-noise standard deviation $\sigma_\omega \sim U(0.05, 0.5)$ pulses/s are drawn as in Appendix A.1, and there is no deviant gain or latency. The observation used for inversion is not $\mathbf{Y}_{\mathrm{std}}$ or $\mathbf{Y}_{\mathrm{dev}}$ individually, but their per-region difference, computed from Eq. (1) under $\mathbf{A}^{(\mathrm{std})} = \mathbf{A}^{(0)}$ and $\mathbf{A}^{(\mathrm{dev})}(t) = \mathbf{A}^{(0)} + g(t)\,\mathbf{B}$, temporally downsampled by a factor of 4, RMS-normalized per case, and passed through $\mathrm{arcsinh}$ for scale compression. This extension is used only to test scalability and observability limits under a much larger candidate set; it is not used for the primary hypothesis-driven results in the main text.

A.2 GENERATOR-LEVEL EFFECT-SIGNATURE ANALYSIS

For each of the 11 candidate effects and $K = 40$ independent draws of the parameters (deviant input gain, deviant latency shift, global coupling, background input, and process-noise scale, sampled from the training distributions in Appendix A.1), we run the Jansen–Rit generator (Appendix A.1) twice with the same random seed: once with only that effect active at $b_e = 0.19$ (the midpoint of the training magnitude range), and once as a null simulation with no active candidate modulation. Because the noise seed is shared, the difference reflects the effect rather than independent noise realizations, and the *effect signature* is the resulting change in the five-source deviant–standard waveform (5×150 samples).

We summarize signatures with two quantities. First, the effect-to-null norm ratio, which we refer to as *SNR*, is the mean Euclidean norm of the effect signature over the 40 draws, divided by a noise floor. The noise floor is the mean, over the same 40 draws, of the Euclidean norm of the difference between two null simulations that differ only in their noise seed (one pair per draw). SNR therefore expresses the size of the effect-induced change relative to the run-to-run variability of the null. Second, the cosine similarity between two effects' mean signatures (averaged over draws) measures how similarly they perturb the observation. Low similarity is desirable, since the two effects can then be told apart; high similarity means they are confusable.

Forward and backward extrinsic effects (Table 4) have similar mean SNR (128.1 and 129.6), whereas intrinsic effects have a mean SNR of 521.9. Signatures of effects terminating at different regions are dissimilar (cosine at most 0.39, mostly near zero), whereas effects terminating at the same region range from 0.40 to 1.00 (an intrinsic effect is assigned its own region as target). In particular, the backward R_IFG→R_STG and forward R_A1→R_STG effects have cosine similarity 0.996. Thus, two connections originating from different regions but terminating at R_STG produce nearly indistinguishable mean observation signatures under the Jansen–Rit generator used to train EC-TCN. This indicates ambiguity in attributing an observed change to one of several incoming connections in this setting; it does not establish non-identifiability for arbitrary observation models.

Per-effect AUPRC is taken from the matched test cohort and averaged over the five EC-TCN seeds. Across the 11 effects, AUPRC is positively associated with SNR (Spearman $\rho = 0.66$, $p = 0.026$). The Pearson correlation between AUPRC and maximum same-target signature similarity ($r = -0.81$, $p = 0.002$) is driven by the near-collinear pair (cosine 0.996), which are also the two lowest-AUPRC effects; the corresponding Spearman correlation is moderate and not significant ($\rho = -0.53$, $p = 0.09$). Given the small number of candidate effects, these associations are reported descriptively.

Table 4: Generator-level effect signatures across 40 parameter draws at $b_e = 0.19$. SNR is the effect-to-null norm ratio defined in the text. The final column gives the most similar effect terminating at the same region.

Effect	Mechanism	SNR	Same-target partner (cosine)
L_A1→L_STG	Forward	130.8	L_STG intrinsic (0.43)
R_A1→R_STG	Forward	134.1	R_IFG→R_STG, backward (1.00)
R_STG→R_IFG	Forward	119.4	R_IFG intrinsic (0.48)
L_STG→L_A1	Backward	130.5	L_A1 intrinsic (0.60)
R_STG→R_A1	Backward	130.5	R_A1 intrinsic (0.60)
R_IFG→R_STG	Backward	127.8	R_A1→R_STG, forward (1.00)
L_A1 intrinsic	Intrinsic	509.7	L_STG→L_A1, backward (0.60)
R_A1 intrinsic	Intrinsic	514.8	R_STG→R_A1, backward (0.60)
L_STG intrinsic	Intrinsic	536.0	L_A1→L_STG, forward (0.43)
R_STG intrinsic	Intrinsic	572.7	R_IFG→R_STG, backward (0.41)
R_IFG intrinsic	Intrinsic	476.3	R_STG→R_IFG, forward (0.48)

A.3 Implementation specifications

The JR EC-TCN implementation follows the architecture and training procedure specified in Section 2.3.1. For completeness, dropout is 0.15 after the first trunk layer, and predicted log variance is clipped to $[-7, 2]$. Training uses AdamW (Loshchilov & Hutter, 2019) with learning rate 2×10^{-3} and weight decay 2×10^{-4}.

For the SPM benchmark, the two conditions and 65 sensors are folded into 130 input channels. The encoder uses Conv1D$(130, 64, 7)$–GELU and Conv1D$(64, 96, 5)$–GELU, followed by adaptive average pooling and a 96-to-128 GELU projection. Separate 11-dimensional heads predict edge presence, effect mean, and log variance; no mechanism-family head is used. The SPM model is trained for 80 epochs with the same optimizer settings specified for JR, but without dropout or gradient clipping. Calibration is fitted only on validation simulations and frozen before evaluation.

SPM benchmark generation. Each of the 11 candidate effects is present with probability 0.38, 12% of cases are forced to be fully null, and a case with no active effect receives one at random. Active magnitudes are $\pm U(0.15, 0.70)$ with a random sign. Responses are generated with the SPM ERP DCM forward model for the five sources (L/R A1, L/R STG, R IFG), observed at 65 sensors and 76 time samples in two conditions. Sensor noise is temporally smoothed Gaussian noise scaled so that its mean channel standard deviation equals that of the noiseless response. There are 600 training, 120 validation, and 200 test cases. DCM is fit independently to each test case with SPM's `spm_dcm_erp` (variational Laplace).

A.4 Additional SPM calibration results

Table 5: Conditional uncertainty calibration on active effects (776 effects) in the 200-case SPM test set. Effects are divided into tertiles of each method's own reported uncertainty scale. Coverage is evaluated against the nominal 95% interval.

	EC-TCN			SPM DCM		
Uncertainty	Coverage	Width	MAE ↓	Coverage	Width	MAE ↓
Low	0.946	0.809	0.174	0.224	0.187	0.272
Middle	0.926	1.411	0.358	0.329	0.550	0.576
High	0.973	2.038	0.411	0.521	1.037	0.603

Table 5 reports calibration conditional on each method's predicted uncertainty for the 200-case SPM test set. Interval width is the mean width of the corresponding 95% intervals.

A.5 Sensitivity of DCM to the number of fitting iterations

A subset of 40 cases was drawn by stratified random sampling from the 200-case test set (12 of the original 60 cases and 28 of the 140 added cases; selection seed 9601) and refit with the maximum number of iterations set to 64, 128, 256, and 512. Observations, template, masks, priors, spatial modes, and scoring were unchanged. We report 95% intervals from 10,000 bootstrap samples at the

case level. Eight single-threaded MATLAB workers ran concurrently, so the s/case values are not comparable with the 79.13 s per case reported in the main text.

Table 6: Effect of the DCM iteration cap on the same 40 SPM test cases. Increasing the cap improves convergence but has little effect on DCM AUPRC. Positive ΔAUPRC favors EC-TCN.

	Recovery		DCM fitting	
Iteration cap	DCM AUPRC	ΔAUPRC [95% CI]	Converged	s/case
64	0.418	0.172 [0.085, 0.256]	7/40 (18%)	89
128	0.415	0.175 [0.087, 0.258]	26/40 (65%)	139
256	0.421	0.169 [0.082, 0.252]	36/40 (90%)	172
512	0.422	0.168 [0.081, 0.251]	40/40 (100%)	183

A.6 Baseline implementations

Method. All estimators are trained and tested on the same SPM benchmark cases and predict the same eleven connectivity targets, but they do not all receive the same input. EC-TCN, random forest, MLP, BiGRU, Transformer, linear/ridge, VAR, and SPM DCM use the 65-sensor observation. The two Sequential Neural Posterior Estimation (SNPE) baselines use a five-source projection instead. This is a linear projection of the 65 sensors onto five source waveforms (the pseudoinverse of the five dominant source topographies), so 65 channels are reduced to five. The input matters because it limits how much of the connectivity can be identified. For random forest, the sensor input gives AUPRC 0.628 (Table 7), whereas the five-source projection gives 0.469 ± 0.002 with the normalized difference waveform and 0.509 ± 0.002 with the standard, deviant, and difference waveforms (five seeds). The SNPE rows are therefore not a like-for-like comparison with the other estimators, and we do not use them to compare inference methods. Method details are as follows:

- *Linear*: logistic regression for presence, ridge regression for magnitude.
- *Random forest*: 250 trees, minimum leaf size 4, maximum features $\sqrt{p}$, classification and regression heads.
- *MLP*: a flattened fully connected network with hidden widths 256 and 128 and dropout 0.15.
- *BiGRU*: a 2-layer bidirectional GRU with 96 hidden units per direction and dropout 0.1.
- *Transformer*: a 2-layer, 4-head encoder with model dimension 64.
- *VAR/Granger* (Granger, 1969; Barnett & Seth, 2014; Shojaie & Fox, 2022): for every source pair, order-5 restricted vs. unrestricted autoregressions; the log ratio of residual variances is fed to the same presence/magnitude read-out. This is a directed predictability statistic, not a coupling estimate.
- *Factorized SNPE* (Cranmer et al., 2020; Lueckmann et al., 2021; Hashemi et al., 2023): an independent 10-component mixture-density posterior for each edge (simulation-based inference), 50 training epochs, 256 posterior samples per case.
- *Joint SNPE/Masked Autoregressive Flow (MAF)* (Cranmer et al., 2020; Lueckmann et al., 2021; Hashemi et al., 2023): one 11-dimensional masked autoregressive flow (5 transform layers, 128 hidden units) over all edges jointly, 100 epochs, 512 posterior samples, 5 seeds.

Both SNPE variants encode the five-source projection with a small multilayer perceptron ($380 \to 128 \to 64$, ReLU) trained jointly with the density estimator, so each baseline learns its own 64-dimensional summary of the observation. We note this is a handicap rather than an advantage at this training-set size: holding seed, split, and test set fixed, replacing the learned encoder with a fixed 64-dimensional linear compression of the same waveform raised factorized SNPE from AUPRC 0.497 to 0.578, and raising the learned encoder's training budget fourfold recovered only to 0.523. With 600 training simulations a jointly trained encoder is data-starved, whereas a closed-form linear fit is not. The SBI numbers reported here are therefore a lower bound on what posterior-density inference achieves on this benchmark. All learned baselines use 5 seeds.

A graph-augmented variant, the effective connectivity temporal graph neural network (EC-TGNN), which is EC-TCN with a graph layer that mixes region embeddings through one fixed adjacency matrix, was trained only on the earlier 60-case version of this benchmark. It reached AUPRC 0.40–0.41 (AUROC 0.53) under identity, shuffled, and structural adjacency and did not pass our acceptance criteria, so it is not compared here.

Results. Table 7 shows that the recovery advantage is not specific to EC-TCN: ridge, random forest, and EC-TCN, all on sensor input, exceed SPM DCM in pooled AUPRC, as does factorized SNPE on the five-source projection.

Table 7: Representative estimator classes on the 200-case SPM benchmark. "Input" is the observation each estimator receives (65 sensors or the five-source projection). Learned-estimator entries are mean ± SD across five retrained seeds. SPM DCM does not use a random generator and therefore has no standard deviation across seeds.

Method class	Representative	Input	AUPRC	Magnitude MAE ↓	Coverage95
Variational inversion	SPM DCM	sensors	0.433	0.484	0.358
Linear discriminative	Ridge	sensors	0.515±0.000	0.318±0.000	**0.945±0.000**
Tree based	Random forest	sensors	**0.628±0.002**	0.362±0.000	0.913±0.009
Temporal neural	EC-TCN (ours)	sensors	0.593±0.005	**0.316±0.009**	0.932±0.007
Posterior density SBI	Factorized SNPE	five-source	0.507±0.015	0.345±0.003	0.692±0.026

The broader comparison supports an estimator-class rather than architecture-specific interpretation. A linear/ridge model reaches AUPRC 0.515, pairwise VAR/Granger reaches 0.495, and random forest attains the highest pooled AUPRC among the learned estimators (0.628 ± 0.002). Table 7 shows the trade-off: random forest has the highest pooled AUPRC, EC-TCN the lowest magnitude error among the representatives shown (0.316 ± 0.009) with coverage 0.932 ± 0.007, and ridge the coverage closest to nominal (0.945). Architecture choice therefore trades ranking accuracy against magnitude recovery and uncertainty calibration, and no single estimator is uniformly best.

The posterior-density SBI results should be read as implementation-specific, for two reasons. First, both SNPE variants learn their own summary of the observation with a small neural encoder trained together with the density estimator. With 600 training simulations this is data-starved. Replacing it with a fixed linear compression of the same waveform raises factorized SNPE AUPRC by about 0.08 (0.497 to 0.578 with seed and split fixed). The joint flow also scores lower than the factorized model (0.394 versus 0.507). Second, the targets are sparse. Each connectivity effect is absent in most simulations and is then exactly zero, so the true prior is a "spike" of probability at zero plus a continuous "slab" of nonzero magnitudes. A flow defines a continuous density and cannot place probability on exactly zero. We therefore add small noise to the zeros during training ($\mathcal{N}(0, 0.005^2)$, clipped to ± 0.015) and call an effect present when the posterior mass lies outside a validation-tuned band around zero. This adds an approximation step that the discriminative estimators, which predict presence directly, do not need. We therefore do not treat these results as evidence against posterior-density SBI in general.

A.7 Scaling experiment and controls

Table 8 reports the full candidate-space scaling results for SparseGraphInverse. The architecture, 68-region observation, and simulation protocol are held fixed while the directed candidate pool increases from $N = 11$ to $N = 1{,}394$. The candidate pool of size N holds the N structurally supported edges with the largest connectome weight, and each case contains 1–10 active edges drawn from the pool. For each N, we report AUPRC, Precision@true-k, active-edge count MAE, and sign accuracy over ground-truth active edges, averaged across three independently trained seeds (6200–6202). These results quantify the progressive loss of edge localization as the candidate space expands while separating localization error from signed-effect recovery. The subsections that follow test alternative explanations for the decline.

Table 8: Candidate-space scaling with SparseGraphInverse. Mean ± SD across three seeds; all 68 regions are observed. Each N uses one simulated cohort: 1,000 training cases for $N \leq 500$ and 3,500 for $N = 1{,}394$. Precision@k uses the true active-edge count k per case; size MAE measures active-edge count error; sign accuracy is evaluated on ground-truth active edges.

N	11	25	50	100	250	500	1,394
AUPRC	.999 ± .000	.811 ± .001	.688 ± .001	.486 ± .002	.260 ± .002	.142 ± .001	.051 ± .001
Prec.@k	.998 ± .000	.759 ± .007	.620 ± .007	.458 ± .006	.278 ± .003	.173 ± .012	.054 ± .005
Size MAE	.11 ± .01	2.09 ± .02	4.50 ± .08	8.94 ± .11	12.32 ± 2.51	16.09 ± 2.72	46.97 ± 20.87
Sign acc.	.999 ± .001	.964 ± .001	.978 ± .000	.970 ± .001	.978 ± .000	.975 ± .001	.973 ± .000

A.7.1 DIRECTLY WIDENED EC-TCN

The candidate-space scaling experiment uses SparseGraphInverse, whose edge scoring is shared across edges, so its size does not depend on N. A reader could therefore suspect that the accuracy decline at large N is specific to this architecture. As a control, we widen the standard EC-TCN so that it scores all N candidate edges directly. The shared trunk (1,170,240 parameters) receives all 68 regional encodings. A presence head, Linear$(128, N)$, outputs one logit per candidate, and a regression head, Linear$(128, 2N)$, outputs a mean and log-variance per candidate. Only these two heads grow with N (Table 9). We omit the family head because the scaling cohorts contain only extrinsic changes. Training used the same optimizer settings as the main model (AdamW, learning rate 2×10^{-3}, weight decay 2×10^{-4}, batch size 64, 100 epochs) with three seeds (6200–6202); $N = 1{,}394$ uses the 3,500-case cohort.

Widened EC-TCN shows the same decline as SparseGraphInverse, and a steeper one: its AUPRC falls from 0.938 ± 0.022 at $N = 11$ to 0.0056 ± 0.0001 at $N = 1{,}394$, against 0.999 to 0.051 for SparseGraphInverse. At $N = 1{,}394$ the base rate is 0.004, so the widened model is close to chance. The decline therefore is not specific to the shared-scoring architecture.

Table 9: Parameter growth for widened EC-TCN compared with SparseGraphInverse. Widened EC-TCN adds 387 parameters per candidate edge; SparseGraphInverse shares its scoring across edges, so its size is fixed.

Estimator	N	Output-head parameters	Total parameters
Widened EC-TCN	11	4,257	1,174,497
Widened EC-TCN	1,394	539,478	1,709,718
SparseGraphInverse	11	shared	9,090
SparseGraphInverse	1,394	shared	9,090

A.7.2 CANDIDATE IDENTITY CONTROL

In the scaling experiment the candidate pools are nested: the pool of size N holds the N edges with the largest connectome weight. Growing N therefore changes which edges are candidates (small pools hold the strongest, shortest edges), not only how many there are. To separate the two, we drew random pools of the same sizes, matched to the full set on connectome-weight decile and on intra- versus interhemispheric class, and checked whether AUPRC still falls as N grows. For each $N \in \{25, 100, 500\}$ we drew three random pools and trained each with three seeds; at $N = 1{,}394$ the pool is the full set, trained with three seeds. It does still fall: AUPRC is 0.897 ± 0.026, 0.585 ± 0.049, 0.141 ± 0.001, and 0.050 ± 0.000 at $N = 25$, 100, 500, and 1,394 (mean $\pm$ SD over pools and seeds). Random pools score higher than nested pools at small N (0.897 versus 0.811 at $N = 25$) but converge at large N (0.141 versus 0.142 at $N = 500$), and the fall from $N = 25$ to $N = 1{,}394$ is similar (about 18-fold versus 16-fold). Candidate identity alone therefore does not explain the scaling curve.

A.7.3 OBSERVABILITY CONTROL

The accuracy decline at large N could simply mean that each case does not contain enough information to localize an edge. To test this, we gave the network more (or less) information at $N = 500$ and checked whether AUPRC changed. The baseline observes all 68 regions for one condition with 150 time samples (AUPRC 0.142). Adding two or four stimulation conditions raises AUPRC by at most 0.002 (95% CIs 0.0002–0.0026 and 0.0009–0.0033), and four times as many temporal samples (600 samples over the same window) lowers it by 0.003. Observing only 34 or 17 regions lowers it by 0.054 or 0.093. More information therefore does not restore localization, while less information makes it worse.

A.7.4 SOURCE COMPETITION

As N grows, each target region has more candidate sources, and the network may be unable to tell which source drove it. To isolate this, we hold the candidate set fixed at $N = 500$ and vary only how many candidate sources compete for each target: 7.35, 14.7, or 25.0. The three settings restrict the 500 edges to 68, 34, or 20 target regions, matched on connectome weight and inter-regional distance. Edge AUPRC falls from 0.137 ± 0.001 to 0.075 ± 0.001 to 0.047 ± 0.000, while target-level AUPRC

stays at 0.993, 0.983, and 0.992. The network finds the affected target but not its source: given the true target, the correct source is ranked first in 0.147, 0.082, and 0.049 of cases, against chance levels of 0.140, 0.075, and 0.047.

A.7.5 DECODER REPLICATION

At $N = 1{,}394$ the AUPRC is low (0.05; base rate 0.004). To check that this reflects the data rather than the decoder, we compared three decoders on a shared temporal encoder. The baseline uses independent edge heads. The hierarchical decoder adds node-involvement, mechanism-family, and active-edge-count supervision that conditions each edge score. The graph decoder additionally passes messages over the candidate adjacency and uses a Hard-Concrete gate for sparse presence probabilities. A linear probe (logistic regression on the RMS amplitude of each edge's source and target regions) provides a model-free reference. All give similar values: one-seed AUPRC is 0.048 for the baseline, 0.047 for both the hierarchical and graph decoders, and 0.045 for the linear probe. These checks are consistent with, but do not by themselves prove, the interpretation that source attribution is what limits localization.

A.8 REAL-EEG RECONSTRUCTION AND SCREENING

Of 68 ds006480 participants, 64 have complete recordings in all four expected conditions. We used the fsaverage template forward model along with the digitized electrode positions of one reference participant and MNE's fsaverage head-to-MRI transform, a three-layer BEM with conductivities $(0.3, 0.006, 0.3)$ S/m, and an ico-4 source space. The dSPM inverse (Dale et al., 2000) in MNE-Python (Gramfort et al., 2013) uses loose orientation 0.2, depth weighting 0.8, fixed diagonal $0.2\,\mu$V sensor noise, and $\lambda^2 = 1/9$. Bilateral transverse temporal and superior temporal cortices and right pars opercularis are extracted by sign-flipped averaging.

Data are average referenced, filtered 0.5–40 Hz, resampled to 250 Hz, epoched from -100 to 500 ms, and baseline corrected over -100 to 0 ms. Epochs are rejected if their amplitude exceeds 200 μV. The passive-control deviant–standard contrast is RMS-normalized, transformed by asinh, and standardized using the frozen model's simulation-derived statistics. This produces the 5×150 input required by EC-TCN without real-data retraining or calibration.

Family-level classifier and DCM. The family-level classifier is trained on a separate cohort of 2,000 training and 300 validation simulations in which each case is assigned uniformly to one of the six mechanism classes. Forward and backward classes modulate four extrinsic connections each with a shared signed magnitude, intrinsic modulation perturbs the self-gain of bilateral A1, and the combined class activates all three (magnitudes $U(0.08, 0.30)$). Ten models are trained independently, and their class probabilities are averaged. SPM DCM was fit with six model families (null, forward, backward, bidirectional, intrinsic, and combined), and families were compared with random-effects Bayesian model selection (Stephan et al., 2009; Penny et al., 2010). For the 42-participant subset, we used a Python port of `spm_BMS_gibbs`, which reproduces SPM's saved output to within 0.001.

A.8.1 EEG SCREENING

The inverse models are trained only on simulations, so their outputs are meaningful only for recordings that resemble that data. We therefore screen every recording before applying a model.

For each recording, we compute 17 features. Twelve are across-region medians and IQRs of six quantities: the autocorrelation at lags 1, 2, 4 and 8, the ratio of high- to low-frequency power, and the temporal standard deviation. The other five are the largest eigenvalues of the inter-regional correlation matrix, normalized by its trace. We estimate the mean and covariance of these features from the training simulations only, using Ledoit–Wolf shrinkage (Ledoit & Wolf, 2004) to keep the covariance estimate stable. We then measure how far each recording lies from that mean with the Mahalanobis distance. A recording is rejected if its distance exceeds the 99th percentile of the validation-simulation distances, and rejected recordings are left out of that model's analyses. This rejects 22 of 64 recordings for the family-level model and 1 of 64 for the edge-level model.

Table 10: Screening of real and surrogate recordings. Entries are the number of recordings retained out of 64. A useful gate retains many real recordings and few surrogates. Surrogates alter temporal or cross-region structure while retaining real-data amplitude information. The whole-brain model uses 68 regions and 50 candidate edges (the strongest by structural connectome weight); it is applied to the same 64 recordings and is not used for the real-EEG results.

Model	Real	Phase-randomized	Shuffled	Noise
Five source (family-level)	42	12	10	10
Whole brain (50 candidate edges)	0	0	0	0

The five-source gate (Table 10) retains 42 real recordings but only 10–12 of each surrogate type, so it separates real data from data whose structure has been destroyed. The whole-brain model's simulations are far from real EEG: it retains none of the 64 real recordings, just as it retains none of the surrogates. Real recordings are still closer to its simulations than any surrogate is (median distance 16, against 24–107 for the surrogates; threshold 7.4), but not close enough to be admitted. We therefore do not interpret that model on real data, and all real-EEG results use the five-source models.

The gate tests similarity of marginal temporal and spectral structure, not whether a recording contains the intended coupling. Phase randomization leaves within-region spectra and autocorrelation invariant; explicit phase-coupling statistics would be required to test that property directly.

A.8.2 Exploratory expanded real-EEG graph

The main real-EEG analysis uses a fixed 11-edge graph of the auditory hierarchy. To check whether that choice limits what we can say about real data, we also analyze a 15-edge graph that adds four interhemispheric connections: L-A1↔R-A1 and L-STG↔R-STG, in both directions (G2). We chose this extension using simulations only. We compared five 15-edge variants (Table 11): G2, G3 (bankssts↔STG, an unobserved region), and three matched-control replicates G4. G2 had the highest validation AUPRC, 0.615 ± 0.002 across three seeds (training seeds 101–103). The base graph G1 scores higher because it has fewer candidates, and G3 and G4 involve a region the network never observes directly, so their validation AUPRC is not a like-for-like ranking with G2. Neither the locked test set nor real EEG played any part in the choice.

We then applied the G2 model to real EEG, after screening recordings with a gate fit to its own training simulations (63 of 64 recordings admitted). Real data have no ground-truth connectivity, so Table 12 reports each edge's mean presence probability, mean signed effect, and sign consistency across the 63 admitted participants (the fraction of participants whose estimated effect has the majority sign; 0.5 is chance). We do not compare this graph with DCM and do not claim that any of these connections is validated.

Table 11: Validation performance of the 11-edge base graph (G1) and five 15-edge variants (three seeds; corrected simulator).

Graph	Added edges	AUPRC	Prec.@k	Size MAE	Sign acc.
G1	base 11-edge graph	0.699 ± 0.009	0.699 ± 0.008	1.50 ± 0.03	0.817 ± 0.008
G2	homologs (L/R A1, L/R STG)	0.615 ± 0.002	0.616 ± 0.009	2.28 ± 0.16	0.774 ± 0.007
G3	bankssts↔STG (unobserved)	0.601 ± 0.005	0.590 ± 0.006	2.87 ± 0.14	0.748 ± 0.015
G4 rep1	temporal pole↔STG (control)	0.596 ± 0.009	0.587 ± 0.020	2.97 ± 0.32	0.753 ± 0.006
G4 rep2	supramarginal↔STG (control)	0.602 ± 0.005	0.598 ± 0.008	2.76 ± 0.23	0.755 ± 0.009
G4 rep3	pars triangularis↔STG (control)	0.599 ± 0.002	0.586 ± 0.001	2.86 ± 0.22	0.751 ± 0.017

A.9 Reproducibility

The reference environment used Python 3.12.12, PyTorch 2.11.0, NumPy 2.4.4, SciPy 1.16.3, scikit-learn 1.8.0, MNE 1.11.0, sbi 0.26.0, nflows 0.14, SPM, and MATLAB R2025b, on an Apple-silicon machine (M4 Pro; MPS and CPU). Primary JR results use five training seeds (9411–9415); SPM training-variability analyses use seeds 9411–9415; candidate-scaling analyses and the widened EC-TCN control use seeds 6200–6202; the 15-edge graph comparison uses training seeds 101–103; the edge-level real-EEG model is a single model (seed 3191); and the real-EEG family analysis uses ten EC-TCN seeds (2070–2079).

Table 12: The 15-edge graph (G2) applied to real EEG: mean presence probability, mean signed effect, and sign consistency across the 63 admitted participants. No ground-truth connectivity exists for these recordings.

Edge	Type	Presence	Mean effect	Sign consistency
L_A1→R_A1	homolog	0.488	−0.040	0.63
R_A1→R_A1	intrinsic	0.502	+0.019	0.51
L_STG→R_STG	homolog	0.399	+0.025	0.56
R_A1→R_STG	forward	0.356	−0.025	0.59
R_STG→R_IFG	forward	0.441	+0.032	0.60
R_STG→R_A1	backward	0.391	−0.074	0.81
R_A1→L_A1	homolog	0.334	+0.036	0.60
L_STG→L_A1	backward	0.326	+0.016	0.54
R_STG→L_STG	homolog	0.332	−0.011	0.56
L_A1→L_STG	forward	0.328	−0.014	0.59
R_IFG→R_STG	backward	0.317	+0.005	0.52
R_STG→R_STG	intrinsic	0.215	−0.024	0.68
R_IFG→R_IFG	intrinsic	0.224	−0.007	0.51
L_A1→L_A1	intrinsic	0.145	−0.041	0.83
L_STG→L_STG	intrinsic	0.189	−0.010	0.63