BOLDFLOW: NETWORK-LEVEL RECOVERY OF FMRI CONNECTIVITY FROM EEG VIA CONDITIONAL FLOW MATCHING

Anonymous authors
Paper under double-blind review

ABSTRACT

Functional connectivity (FC) analyses operate on correlations between regional blood-oxygen-level-dependent (BOLD) signals, but estimating them requires functional magnetic resonance imaging (fMRI). Inferring fMRI-like connectivity from scalp electroencephalography (EEG) could broaden access, yet per-region regressors and joint mean-squared-error decoders collapse toward the conditional mean, discarding the residual cross-region covariance. We present BOLDFlow, which models the conditional distribution of parcellated BOLD activity using conditional flow matching with a learned EEG-conditioned source. Sampled trajectories retain the model's residual cross-region covariance, while repeated draws provide input-conditional ensemble spread without a separate uncertainty head. In subject-disjoint evaluations on two simultaneous in-scanner EEG–fMRI datasets spanning rest and sleep, BOLDFlow improves measured-FC recovery (NeuroBOLT: 0.428 to 0.584; OpenNeuroSleep: 0.340 to 0.528) and temporal correlation (0.272 to 0.326 and 0.151 to 0.212, respectively), and scales to output parcellations with 512 components. The recovered connectomes retain individual-specific and community structure, while ensemble spread attains nominal coverage after single-parameter recalibration.

1 INTRODUCTION

Functional magnetic resonance imaging (fMRI) remains the dominant non-invasive window into spatially resolved brain dynamics (Logothetis et al., 2001), yet scanner cost, motion constraints, and incompatibility with some implants limit access to this view of brain activity. These constraints motivate recovering fMRI-related information from electroencephalography (EEG), an inexpensive, portable complement with millisecond temporal resolution. Decades of simultaneous EEG-fMRI studies have established structured, region-specific correlations between EEG spectral power and the blood-oxygen-level-dependent (BOLD) signal (Goldman et al., 2002; Laufs et al., 2003). These relationships are consistent with BOLD's sensitivity to local field potentials (Logothetis et al., 2001) and with the distinctive electrophysiological signatures of canonical resting-state networks (Mantini et al., 2007). These findings support predicting regional BOLD from EEG to recover measured-fMRI connectivity.

Clinical and cognitive neuroimaging increasingly examines brain function through networks of interacting regions, characterized by functional connectivity (FC) between regional BOLD time courses. FC patterns stratify depression into neurophysiological biotypes (Drysdale et al., 2017), identify individuals across sessions (Finn et al., 2015), and organise cognition and behaviour at population scale (Smith et al., 2015). Neurodegenerative, psychiatric, and neuro-oncological conditions are likewise studied as *network disorders* (Xia et al., 2018; Fornito et al., 2015). For EEG-derived predictions to support these analyses, they must preserve the relationships between regional signals as well as their individual time courses. Predicting parcellated BOLD places the reconstructed signals in the same regional representation used for FC estimation, regional statistics, and community analysis.

Regional EEG-to-fMRI methods include separate regressors for individual targets (Meir-Hasson et al., 2014; Li et al., 2024). Under squared-error training, these regressors and joint decoders share the conditional mean as their population optimum. Such mean predictions suppress residual cross-ROI

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.

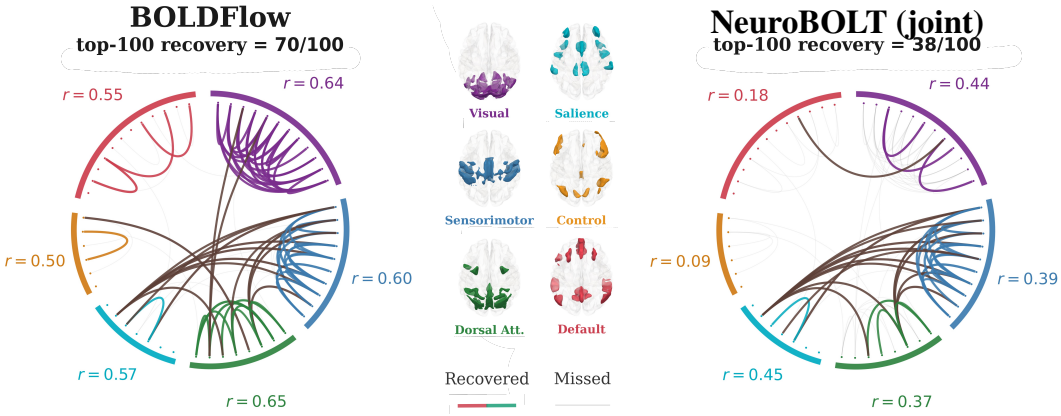

Figure 1: **Functional-connectivity recovery: BOLDFlow vs. NeuroBOLT (joint-output adaptation of Li et al., 2024).** Connectograms show the $K{=}100$ strongest ground-truth connections ($\approx 5\%$ edge density) on the full 64-component Fisher-z group-average FC matrix (1) over $n{=}29$ held-out NeuroBOLT scans (5-fold CV). DiFuMo-64 components are grouped by the per-component network labels in the atlas metadata (Dadi et al., 2020) into six cortical networks. Coloured chords lie in the method's own top-K (*recovered*); grey chords are *missed*. Per-network r is the Pearson correlation between upper-triangular entries of the group-average ground-truth and predicted FC, restricted to edges with at least one endpoint in the network. Full protocol in Appendix F.

covariance (Prop. 1, Section 3), creating a mismatch between a pointwise reconstruction objective and the network relationships we seek to recover. The information available to the predictor also depends on its temporal context: typical configurations combine fewer than 100 regions of interest (ROIs) with short EEG windows, including 16-second windows in NeuroBOLT (Li et al., 2024), whereas longer contexts can capture additional hemodynamic information (Glover, 1999). Longer EEG contexts and joint distributional prediction address these temporal and spatial aspects of connectivity recovery.

We introduce BOLDFlow,[1] which models the conditional distribution of regional BOLD activity given EEG. BOLDFlow encodes EEG using REVE (El Ouahidi et al., 2025), an EEG foundation model pretrained on over 60,000 hours of EEG, together with a multi-scale spectral module. The resulting representation conditions a flow decoder trained with conditional flow matching (CFM) (Lipman et al., 2023; Tong et al., 2024) to transport samples from a learned EEG-conditioned Gaussian source supervised under β-NLL (Seitzer et al., 2022). Jointly sampled regional trajectories carry the model's mean and residual covariance into FC estimation, whereas averaging those trajectories progressively attenuates the residual contribution (Proposition 1 and Corollary 1). Repeated draws from the same model also provide input-conditional ensemble spread for uncertainty quantification (UQ).

On two simultaneous EEG–fMRI datasets spanning rest and sleep, BOLDFlow improves FC recovery and temporal reconstruction over the compared baselines. Our sampling analysis shows that individual sampled trajectories agree more closely with measured FC than ensemble-mean predictions from the same trained model. Constant-input, subject-pairing, and temporal-alignment controls assess the dependence of FC recovery on the supplied EEG beyond shared population structure. We also evaluate ensemble spread through calibration, error ranking, and FC-edge selection, with comparisons to last-layer Laplace (Daxberger et al., 2021) and an error predictor (Bramlage & Curio, 2025).

Our contributions are:

1. **Joint BOLD prediction for connectivity recovery.** We introduce BOLDFlow, a conditional flow-matching model with a learned EEG-conditioned source that generates regional BOLD trajectories jointly for FC estimation.
2. **Covariance analysis of connectivity readouts.** We characterize residual covariance loss under conditional-mean prediction and ensemble averaging. At a fixed checkpoint, sampled trajectories yield higher FC agreement than ensemble-mean predictions.

[1] Code: anonymous.4open.science/r/boldflow-3FC6.

3. **Network recovery across rest and sleep.** BOLDFlow raises FC Corr from 0.428 to $\mathbf{0.584}$ on NeuroBOLT and from 0.340 to $\mathbf{0.528}$ on OpenNeuroSleep, alongside improved temporal reconstruction. Input controls and community analyses assess EEG specificity and recovery of network structure.

2 Related Work

EEG-to-fMRI synthesis. Early computational approaches used hand-crafted EEG features and region-specific regressors (Meir-Hasson et al., 2014), followed by CNN-based neural transcoding (Liu & Sajda, 2019) and attention over electrode relationships for volumetric synthesis (Calhas & Henriques, 2023). A strong parcellated baseline is NeuroBOLT (Li et al., 2024), which combines a pretrained LaBraM encoder (Jiang et al., 2024) with multi-scale spectral features for resting-state synthesis. Alongside these regional models, NT-ViT (Lanzino et al., 2025) reconstructs fMRI volumes, while CATD (Yao et al., 2025) generates cortical surface maps and evaluates both spatial and temporal fidelity. The parcellated baselines motivate our comparison of pointwise regression with distributional decoding, and of the standard 16-second context with longer EEG windows.

Conditional flow matching. Flow matching (Lipman et al., 2023) provides a simulation-free framework for training continuous normalising flows, with efficiency improvements via mini-batch optimal-transport couplings (Tong et al., 2024; Pooladian et al., 2023). AdaLN-Zero conditioning from the Diffusion Transformer (Peebles & Xie, 2023) and its feature-wise-modulation precursors (Perez et al., 2018) are our conditioning template. Flow matching has been applied to biomolecular and physiological signals (Bose et al., 2024; Bondar et al., 2025). Condition-specific anchors (Issachar et al., 2025) and Gaussian-process sources (Kollovieh et al., 2025) establish informative source distributions as a generative-modeling design choice. BOLDFlow studies an EEG-conditioned diagonal Gaussian source with auxiliary β-NLL supervision and joint regional transport for connectivity recovery. Its evaluation connects source and velocity ablations to a fixed-checkpoint comparison of sampled-trajectory and ensemble-mean FC. This comparison examines how the readout exposes the learned distribution's covariance, while input controls assess EEG specificity and repeated draws support uncertainty estimation.

Uncertainty quantification. Most EEG-to-fMRI methods provide only point predictions. UQ methods such as MC-Dropout (Gal & Ghahramani, 2016) and deep ensembles (Lakshminarayanan et al., 2017) provide comparison points, while heteroscedastic output heads can alter the mean through mean-variance coupling (Stirn et al., 2023). BOLDFlow's distributional prior, the same architectural element that supplies the sampled trajectories used for FC recovery (Section 3), regularizes the source toward input-conditional target moments, supervised by β-NLL (Seitzer et al., 2022) which modifies variance-dependent loss weighting; multi-sample inference reuses this element as a step toward native UQ (Section 4). An extended discussion of all three lineages, with additional citations and historical context, is in Appendix H.

3 Method

Problem Setup. Let $\mathbf{X} \in \mathbb{R}^{C\times N}$ be the preprocessed EEG recording and $\{\mathbf{y}_t\}_{t=1}^{L}$ the synchronized fMRI sequence, with each volume represented by R DiFuMo components (Dadi et al., 2020). For anchor TR t, the model receives $\mathbf{x}_t = \mathbf{X}[:, n_t - T_{\mathrm{ctx}} : n_t]$, where n_t is the aligned EEG sample and $\Delta_{\mathrm{ctx}} = T_{\mathrm{ctx}}/f_s$ is the context duration. This window is chosen to cover the temporal support of the hemodynamic response (Glover, 1999).

The target depends on the setting: *(i) sequence-to-one* predicts $\mathbf{y}_t \in \mathbb{R}^R$; *(ii) sequence-to-sequence* predicts T_{out} consecutive volumes ending at t, $\mathbf{y}_{t-T_{\mathrm{out}}+1:t} \in \mathbb{R}^{T_{\mathrm{out}}\times R}$, flattened to dimension $d = T_{\mathrm{out}}R$. We write $(\mathbf{x}, \mathbf{y})$ for a matched training pair and model $p(\mathbf{y} \mid \mathbf{x})$ to permit repeated stochastic sampling.

Functional Connectivity and Conditional-Mean Prediction. Functional connectivity (FC) is computed over the full parcellated time series $\{\mathbf{y}_t\}_{t=1}^{L}$. Let $\bar{\mathbf{y}} = L^{-1}\sum_{t=1}^{L}\mathbf{y}_t$ denote the temporal mean. The empirical BOLD covariance and its correlation-normalized FC matrix (Friston, 1994) are

defined as

$$\widehat{\Sigma}_{\mathrm{BOLD}} = \frac{1}{L-1}\sum_{t=1}^{L}(\mathbf{y}_t - \bar{\mathbf{y}})(\mathbf{y}_t - \bar{\mathbf{y}})^\top, \qquad \widehat{\mathbf{F}} = \mathrm{diag}(\widehat{\Sigma}_{\mathrm{BOLD}})^{-1/2}\widehat{\Sigma}_{\mathrm{BOLD}}\,\mathrm{diag}(\widehat{\Sigma}_{\mathrm{BOLD}})^{-1/2}. \tag{1}$$

For an R-dimensional BOLD vector at a generic TR, let $\mathcal{E}$ denote the EEG information available to its predictor. Under the joint distribution relevant to a scan, the law of total covariance gives

$$\Sigma_{\mathbf{y}} = \mathbb{E}_{\mathcal{E}}\left[\mathrm{Cov}(\mathbf{y} \mid \mathcal{E})\right] + \mathrm{Cov}_{\mathcal{E}}\left(\mathbb{E}[\mathbf{y} \mid \mathcal{E}]\right) = A + B. \tag{2}$$

Here, B is the covariance of the conditional mean and A is the expected conditional residual covariance. Proposition 1 identifies the covariance retained by the population MSE optimum and its corresponding FC.

Proposition 1 (Covariance of MSE-optimal point predictors). *Let $\mathbb{E}\|\mathbf{y}\|^2 < \infty$ and let $f^\star$ minimize population MSE over square-integrable predictors. With A, B as in (2), define the correlation-normalised forms $\tilde{B} = \mathrm{diag}(B)^{-1/2}B\,\mathrm{diag}(B)^{-1/2}$ and $\tilde{C} = \mathrm{diag}(A+B)^{-1/2}(A+B)\,\mathrm{diag}(A+B)^{-1/2}$. Assume positive diagonal entries in B and non-zero variance in each upper-triangle vector. Then:*

(i) *Any MSE-optimal point predictor satisfies $f^\star(\mathcal{E}) = \mathbb{E}[\mathbf{y} \mid \mathcal{E}]$ (Bishop, 2006) and $\mathrm{Cov}(f^\star(\mathcal{E})) = B$, so A is absent from the predictor's covariance.*

(ii) *Its FC is $\tilde{B}$, with similarity to measured-BOLD population FC given by $\rho_{\mathrm{CM}} := \mathrm{Corr}(\mathrm{vec}_\triangle\,\tilde{B},\ \mathrm{vec}_\triangle\,\tilde{C})$. This similarity equals one exactly when the two upper-triangle vectors are related by a positive affine transformation.*

The proof is given in Appendix G. Per-ROI and joint MSE regression share the conditional mean as their population target. For connectivity recovery, BOLDFlow models $p(\mathbf{y} \mid \mathbf{x})$ and constructs regional trajectories by joint sampling. The learned source supplies input-dependent starting moments, the velocity field couples regional outputs, and the trajectory readout retains the model's residual variation. The next sections specify these three parts of the method; FC is evaluated within each held-out scan.

Learned-Source Conditional Flow Matching. To model $p(\mathbf{y} \mid \mathbf{x})$, we adopt the flow-matching framework (Lipman et al., 2023). An encoder maps the EEG window to a conditioning vector $\mathbf{c} = \mathrm{Enc}(\mathbf{x})$, and a prior head then produces the parameters of an input-dependent Gaussian source

$$q_\phi(\mathbf{z}_0 \mid \mathbf{c}) = \mathcal{N}\left(\boldsymbol{\mu}_\phi(\mathbf{c}), \mathrm{diag}\,\boldsymbol{\sigma}^2_\phi(\mathbf{c})\right). \tag{3}$$

During training, we draw $\mathbf{z}_0 \sim q_\phi(\cdot \mid \mathbf{c})$ and a flow time $\tau \sim \mathcal{U}[0,1]$, then define the straight-line probability path $\mathbf{z}_\tau = (1-\tau)\mathbf{z}_0 + \tau\mathbf{y}$ with velocity $\mathbf{u}_\tau = \mathbf{y} - \mathbf{z}_0$. The velocity network $\mathbf{v}_\theta$ is trained via the conditional flow-matching (CFM) objective

$$\mathcal{L}_{\mathrm{CFM}}(\Omega) = \mathbb{E}\left[\|\mathbf{v}_\theta(\mathbf{z}_\tau, \tau \mid \mathbf{c}) - \mathbf{u}_\tau\|_2^2\right]. \tag{4}$$

At inference, we sample $\mathbf{z}_0$ and integrate the ODE $d\mathbf{z}_\tau/d\tau = \mathbf{v}_\theta(\mathbf{z}_\tau, \tau \mid \mathbf{c})$ over $\tau \in [0,1]$ using 50 Euler steps. The terminal state $\mathbf{z}_1$ gives a stochastic BOLD prediction $\hat{\mathbf{y}}$, and different source draws for the same $\mathbf{x}$ define the model's predictive distribution $\hat{p}_\Omega(\cdot \mid \mathbf{x})$, where Ω collects the encoder, prior, and velocity parameters.

Training Objectives and Source Anchoring. To regularize the source mean and diagonal covariance toward input-conditional target moments, we use the β-NLL objective (Seitzer et al., 2022). For target dimension d and $\beta = 0.5$, the prior objective is

$$\mathcal{L}_{\mathrm{prior}} = \mathbb{E}\left[\frac{1}{2}\sum_{j=1}^{d}\mathrm{sg}\left(\sigma^{2\beta}_{\phi,j}(\mathbf{c})\right)\left(\frac{(y_j - \mu_{\phi,j}(\mathbf{c}))^2}{\sigma^2_{\phi,j}(\mathbf{c})} + \log\sigma^2_{\phi,j}(\mathbf{c})\right)\right]. \tag{5}$$

Here, $\mathrm{sg}(\cdot)$ stops gradients through the variance-dependent weight while the Gaussian NLL term updates the mean and variance. In isolation and under sufficient capacity, the estimating equations

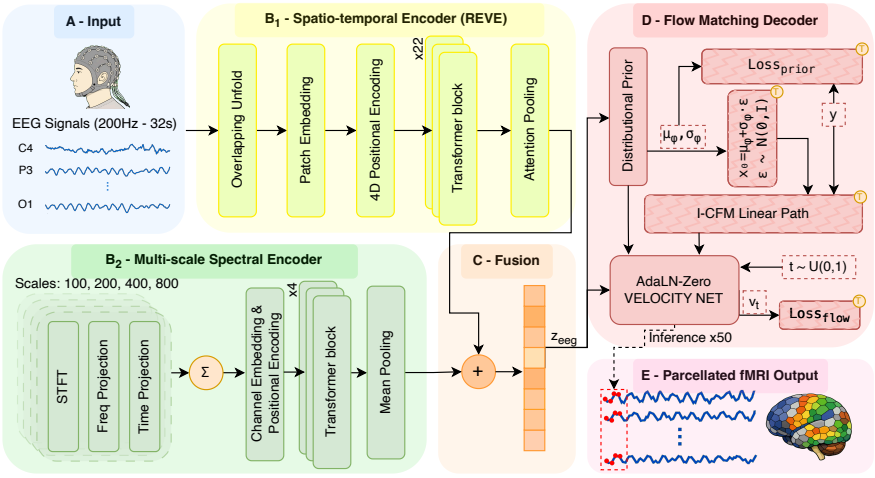

Figure 2: **BOLDFlow architecture.** (A) A scalp-EEG context window is extracted for each target fMRI TR. (B_1) A REVE-based spatio-temporal encoder captures EEG spatial and temporal structure. (B_2) A NeuroBOLT-style multi-scale spectral encoder captures frequency-domain structure. (C) Additive fusion produces the EEG embedding that conditions both the source and velocity field. (D) An AdaLN-Zero velocity field transports draws from the EEG-conditioned Gaussian source to BOLD outputs using 50 Euler steps per draw. Training jointly minimizes the CFM and β-NLL losses. (E) Predicted DiFuMo BOLD blocks are overlap-averaged to form the trajectory used for FC estimation.

target $\boldsymbol{\mu}_\phi(\mathbf{c}) = \mathbb{E}[\mathbf{y} \mid \mathbf{c}]$ and $\boldsymbol{\sigma}^2_\phi(\mathbf{c}) = \mathbb{E}[(\mathbf{y} - \boldsymbol{\mu}_\phi(\mathbf{c}))^2 \mid \mathbf{c}]$. During joint training, the flow objective also updates the source through reparameterized samples, aligning source learning with the transport task. The joint velocity field produces cross-ROI dependencies during integration.

The full training objective is $\mathcal{L}(\Omega) = \mathcal{L}_{\text{CFM}} + \lambda \mathcal{L}_{\text{prior}}$.

Functional Connectivity from Stochastic Predictions. To construct a sampled trajectory, we draw one output block per anchor and average the predictions covering each TR, as defined in Eq. (8). Writing $\mathcal{E}$ for the collection of EEG windows contributing to a given TR, the covariance of its aggregated prediction decomposes as

$$\Sigma_\Omega = \mathbb{E}_\mathcal{E}[\text{Cov}_\Omega(\hat{\mathbf{y}} \mid \mathcal{E})] + \text{Cov}_\mathcal{E}(\mathbb{E}_\Omega[\hat{\mathbf{y}} \mid \mathcal{E}]) = A_\Omega + B_\Omega.$$

Computing FC from this trajectory gives *sampled-trajectory FC*. Repeating the full procedure independently M times and averaging the resulting trajectories before computing FC gives *ensemble-mean FC*.

Corollary 1 (Sampled-trajectory FC vs ensemble-mean FC). *A sampled prediction has covariance $A_\Omega + B_\Omega$. For M conditionally independent repetitions, the ensemble-mean covariance is*

$$\text{Cov}(\bar{\mathbf{y}}_{t,M}) = B_\Omega + M^{-1} A_\Omega. \tag{6}$$

Under stationarity, ergodicity, and finite second moments, empirical trajectory covariances converge to these model-side quantities. If the aggregated predictions match the data's conditional first and second moments given the same $\mathcal{E}$, then $(A_\Omega, B_\Omega) = (A, B)$ and the sampled predictor recovers the data-side covariance.

The proof is given in Appendix G. BOLDFlow reports FC from a sampled trajectory. Repeated trajectories additionally provide pointwise spread for UQ; their mean defines a separate prediction whose residual covariance scales with $1/M$. At fixed checkpoints, trajectory averaging improves pointwise accuracy while reducing FC recovery (Appendix G). Regional prediction intervals use the ensemble mean as their center and the spread as their width estimate (Appendix E). Algorithm 1 summarizes training and both trajectory readouts.

Architecture. Figure 2 summarizes the architecture. Dimensions and hyperparameters are given in Appendix B.

Spatio-temporal stream. The spatio-temporal encoder follows REVE (El Ouahidi et al., 2025), pretrained on more than $60{,}000$ hours from 92 datasets. It processes overlapping EEG patches with a 4D Fourier-coordinate positional encoding (Tancik et al., 2020) of the channel's 3D scalp coordinates and patch time. A transformer stack and attention pooling produce a representation $\mathbf{h}_{\text{temp}} \in \mathbb{R}^{512}$.

Spectral stream. Following NeuroBOLT (Li et al., 2024), the spectral stream computes multi-scale STFT magnitudes, projects each scale, sums the resulting multi-resolution spectral embeddings, tokenises by channel, and processes the tokens using a linear-attention transformer (Katharopoulos et al., 2020). Mean pooling gives a spectral representation $\mathbf{h}_{\text{spec}} \in \mathbb{R}^{512}$.

Fusion. The two streams are combined by additive fusion, $\mathbf{c} = \text{GELU}(\mathbf{h}_{\text{temp}} + \mathbf{h}_{\text{spec}})$, at all evaluated parcellations. Appendix C compares this choice with gated cross-attention fusion.

Decoder. The prior head uses a two-layer MLP trunk with separate mean and scale heads. The velocity field is a residual MLP with four AdaLN-Zero blocks (Peebles & Xie, 2023). EEG and sinusoidal flow-time embeddings each pass through a two-layer MLP; their sum determines each block's scale, shift, and residual gate in Eq. (7). Zero-initialized modulation starts each block at the identity.

Output and aggregation. Sequence-to-one returns one BOLD vector; sequence-to-sequence outputs are reshaped into T_{out} volumes ending at the anchor. Each anchor uses an independent source draw, and the overlapping blocks are averaged into the trajectory defined above. With $T_{\text{out}} = 4$, this offline reconstruction can use EEG up to three TRs after a target volume. Appendix C analyzes how block length and EEG context affect the aggregated prediction.

4 Experiments

Datasets and baselines. We evaluate BOLDFlow on two simultaneous EEG–fMRI datasets spanning rest and sleep. **NeuroBOLT** (Li et al., 2024) contains 29 resting-state scans from 22 healthy subjects, with 26 EEG channels and fMRI acquired on a 3T Philips scanner. For **OpenNeuroSleep** (Gu et al., 2023), released as OpenNeuro ds003768, we evaluate 229 rest/sleep scans from 33 healthy subjects using 30 EEG channels, 22 shared with NeuroBOLT. Both datasets use fMRI TR=2.1 s and EEG resampled to 200 Hz, while differing in scanner, acquisition protocol, and montage.

The comparisons include our joint-output MSE adaptation of **NeuroBOLT** and its original per-ROI formulation, denoted **NeuroBOLT**†. To examine the contributions of the encoder and decoder, we also evaluate a NeuroBOLT-encoder/BOLDFlow-decoder hybrid (**NB enc. w/ BF dec.**) and MLP readouts on frozen or fine-tuned REVE embeddings (**REVE-NoFT/REVE-FT**). Preprocessing and baseline implementations are detailed in Appendices A and B.

Metrics. We use 5-fold subject-disjoint cross-validation with 3 random seeds per fold, for 15 runs in total, and report standard deviations across runs. Prediction accuracy is evaluated with MSE and T. Corr, while functional-connectivity recovery is evaluated with FC Corr. For measured and predicted regional time courses $\mathbf{y}_{r,1:L}$ and $\hat{\mathbf{y}}_{r,1:L}$, with FC matrices computed by Eq. (1), we report

$$\text{T. Corr} = \frac{1}{R}\sum_{r=1}^{R} \rho(\mathbf{y}_{r,1:L}, \hat{\mathbf{y}}_{r,1:L}), \qquad \text{FC Corr} = \rho\left(\text{vec}_{\triangle}(\widehat{\mathbf{F}}_{\text{pred},S}), \text{vec}_{\triangle}(\widehat{\mathbf{F}}_{\text{data},S})\right).$$

Here ρ denotes Pearson correlation, S is the evaluated component set, and $\text{vec}_{\triangle}(\cdot)$ vectorizes the off-diagonal upper-triangular entries of a symmetric matrix. T. Corr measures temporal reconstruction within parcels, whereas FC Corr measures recovery of the between-parcel correlation structure. For DiFuMo-64, MSE and T. Corr summarize all 64 outputs; the primary FC Corr is computed on the fixed 55-component cortical mask, applied identically to predictions and targets. FC is evaluated within scans, with group-level visualizations used to summarize network structure.

BOLDFlow's main-comparison MSE, T. Corr, and FC Corr use the same sampled trajectory ($M = 1$, Section 3). Native ensemble UQ uses $M = 50$ repetitions to estimate regional BOLD means and pointwise spread. We evaluate the resulting prediction intervals through calibration error and Cov@95%, and uncertainty scores through their Spearman correlation with absolute residuals (Appendix E).

Table 1: **Cross-method comparison.** Five-fold subject-disjoint CV with three seeds per fold (mean $\pm$ standard deviation across 15 runs). MSE and T. Corr summarize the 64 output components; the primary DiFuMo-64 FC Corr uses the fixed 55-component cortical evaluation mask. NeuroBOLT (joint) is our joint-output adaptation of the original per-region method (Li et al., 2024); NeuroBOLT† is the per-ROI comparison. The hybrid uses the NeuroBOLT encoder with the BOLDFlow decoder; REVE-NoFT/-FT use MLP readouts on frozen/fine-tuned REVE. Best in **bold**.

Method	NEUROBOLT			OPENNEUROSLEEP		
	MSE ↓	T. Corr. ↑	FC Corr. ↑	MSE ↓	T. Corr. ↑	FC Corr. ↑
BOLDFlow (ours)	**0.239**$_{\pm 0.002}$	**0.326**$_{\pm 0.008}$	**0.584**$_{\pm 0.017}$	**0.255**$_{\pm 0.001}$	**0.212**$_{\pm 0.001}$	**0.528**$_{\pm 0.017}$
NB enc. w/ BF dec.	0.245$_{\pm 0.002}$	0.295$_{\pm 0.012}$	0.550$_{\pm 0.017}$	0.257$_{\pm 0.001}$	0.190$_{\pm 0.008}$	0.489$_{\pm 0.020}$
NeuroBOLT (joint)	0.246$_{\pm 0.002}$	0.272$_{\pm 0.013}$	0.428$_{\pm 0.014}$	0.260$_{\pm 0.001}$	0.151$_{\pm 0.007}$	0.340$_{\pm 0.021}$
NeuroBOLT†	0.242$_{\pm 0.002}$	0.284$_{\pm 0.013}$	0.253$_{\pm 0.021}$	0.406$_{\pm 0.002}$	0.132$_{\pm 0.005}$	0.232$_{\pm 0.020}$
REVE-NoFT	0.256$_{\pm 0.001}$	0.095$_{\pm 0.004}$	0.266$_{\pm 0.016}$	0.265$_{\pm 0.001}$	0.038$_{\pm 0.002}$	0.157$_{\pm 0.015}$
REVE-FT	0.243$_{\pm 0.001}$	0.251$_{\pm 0.011}$	0.354$_{\pm 0.016}$	0.262$_{\pm 0.001}$	0.122$_{\pm 0.003}$	0.250$_{\pm 0.025}$

Implementation. After 0.5 Hz high-pass filtering and per-channel z-scoring, EEG is processed in 32 s windows ending at each fMRI trigger. We use additive fusion, four-volume seq2seq outputs, and I-CFM training with objective $\mathcal{L}_{\mathrm{CFM}} + \lambda\mathcal{L}_{\mathrm{prior}}$, setting λ=1 and β=0.5 (Seitzer et al., 2022). The model has 96.4M parameters, including 70.2M from pretrained REVE, and each training run takes approximately 2.5 h on an A100 40 GB GPU. Each NeuroBOLT fold reserves three validation subjects for checkpoint selection. The final configuration was developed through approximately 300 NeuroBOLT runs and then trained separately on OpenNeuroSleep, which was excluded from configuration selection. Full optimization and development details are in Appendices A and B.

Connectivity recovery across datasets BOLDFlow improves connectivity recovery together with regional reconstruction, leading the compared methods on all three metrics in both datasets (Table 1). Relative to joint-output NeuroBOLT, FC Corr rises from 0.428 to $\mathbf{0.584}$ on NEUROBOLT and from 0.340 to $\mathbf{0.528}$ on OPENNEUROSLEEP; T. Corr increases from 0.272 to 0.326 and from 0.151 to 0.212, respectively, alongside lower MSE. The configuration selected on NeuroBOLT therefore also improves recovery when trained on the rest/sleep dataset.

The group-level connectogram in Fig. 1 shows how these aggregate gains translate into recovered connections: BOLDFlow retains $70/100$ of the strongest measured edges, versus NeuroBOLT's $38/100$. Similarity spans $r \in [0.50, 0.65]$ across all six cortical networks, including Control and Default, where NeuroBOLT reaches 0.09 and 0.18. The improvement thus extends across networks with different baseline recovery levels, preserving more of the strong connections that organize the measured connectome.

Downstream FC validation. Beyond edge agreement, subject fingerprinting across predicted scan halves reaches $\sim 18.5\times$ chance for BOLDFlow, versus $\sim 16\times$ for joint-MSE NeuroBOLT and $\sim 15.5\times$ for its per-ROI variant (Appendix F). Louvain community recovery also improves: agreement with measured partitions, assessed by adjusted Rand index, is $\sim 2.6\times$ that of joint-MSE and $\sim 4.4\times$ that of per-ROI NeuroBOLT (Appendix F). These results connect improved FC recovery to more stable individual signatures and closer agreement with measured network organization. Adapting the decoder on one recording further improves FC recovery on the same subject's other recording in all 14 adaptation/test directions across seven subjects, with a subject-clustered interval of $[0.077, 0.293]$ for the FC Corr gain (Appendix F).

EEG-conditioned recovery. Retraining the same architecture with constant EEG lowers FC Corr from 0.584 to 0.264, showing the benefit of conditioning on the recording. After removing the training-population FC template, correct-subject EEG outperforms wrong-subject EEG in all five folds, with a 95% interval of $[0.070, 0.280]$ for the matched-minus-wrong contrast. Within recordings, aligned EEG also improves dynamic-FC recovery over a 180 s displacement, with an interval of $[0.016, 0.041]$ for the aligned-minus-displaced contrast. This comparison holds subject identity fixed, extending the evidence from subject pairing to temporal alignment. Together, these controls link recovery to the availability, identity, and timing of the supplied EEG (Appendix D).

Table 2: **Uncertainty quantification on NeuroBOLT** (DiFuMo-64; 5 folds, 3 seeds; mean ± std over 15 runs). Top: fixed-σ pipeline (T. Corr 0.321; heteroscedastic head $0.276^{\dagger}$). Middle/bottom: post-hoc/native UQ on BOLDFlow (reference sampled-trajectory T. Corr 0.326; native intervals are centered on ensemble-mean regional BOLD, Appendix E). **Bold**: best per column; Cov@95% closest to 0.95 is best.

Method	Spearman ↑	Cal. Err. ↓	Cov@95%
Source noise (σ=0.1, K=20)	$0.034_{\pm 0.002}$	$0.500_{\pm 0.004}$	$0.038_{\pm 0.001}$
MC-Dropout (p=0.1, K=20)	$0.005_{\pm 0.002}$	$0.520_{\pm 0.003}$	$0.012_{\pm 0.001}$
Deep ensembles (M=5 seeds)	$0.045_{\pm 0.004}$	$0.470_{\pm 0.004}$	$0.080_{\pm 0.003}$
Heteroscedastic output head†	$0.060_{\pm 0.004}$	$0.245_{\pm 0.004}$	$0.620_{\pm 0.005}$
Error Predictor (EP)	$0.110_{\pm 0.006}$	$0.095_{\pm 0.001}$	$0.842_{\pm 0.002}$
Last-layer Laplace	$0.195_{\pm 0.009}$	$0.305_{\pm 0.002}$	$0.460_{\pm 0.006}$
EP + Laplace rank-avg fusion	$0.175_{\pm 0.006}$	$0.025_{\pm 0.001}$	$0.882_{\pm 0.001}$
Raw ensemble	$0.155_{\pm 0.006}$	$0.330_{\pm 0.003}$	$0.385_{\pm 0.004}$
+ scalar recalibration ($\alpha \approx 4.6$)	$0.155_{\pm 0.006}$	$\mathbf{0.011}_{\pm 0.001}$	$\mathbf{0.948}_{\pm 0.002}$
+ rank-avg fusion with Laplace	$\mathbf{0.225}_{\pm 0.006}$	$0.020_{\pm 0.001}$	$0.892_{\pm 0.002}$

Robustness across parcellations. The FC advantage persists at 256 and 512 output components, extending recovery to finer regional representations with the same EEG montage (Appendix C). This comparison varies the granularity of the BOLD targets while keeping the acquired EEG information fixed.

Where the gain comes from We first examine how the gains depend on the EEG representation and decoder, then assess the decoder's architecture and temporal context.

First axis (decoder vs encoder). The hybrid in Table 1 examines how much of the improvement is retained with the baseline encoder. Combining the NeuroBOLT encoder with BOLDFlow's decoder raises FC Corr to 0.550 on NeuroBOLT and 0.489 on OpenNeuroSleep, recovering much of the full model's improvement. The REVE-based representation adds a further 0.034/0.039 in FC Corr and improves temporal reconstruction on both datasets. The gains with the baseline encoder motivate a closer comparison of the decoder configurations.

Second axis (within-decoder isolation). Proposition 1 motivates evaluating decoder choices beyond pointwise reconstruction: at the population MSE optimum, predictions omit the conditional residual covariance that contributes to functional connectivity. We therefore hold REVE and the spectral branch fixed and compare regression with flow-based decoding (Table 6, Appendix C). Gaussian-source I-CFM raises FC Corr from MLP regression's 0.420 to 0.461, while T. Corr falls from 0.315 to 0.281 and MSE rises from 0.241 to 0.252. OT coupling reaches FC Corr 0.470, also at the cost of pointwise accuracy. Full BOLDFlow combines a learned source, β-NLL supervision, and I-CFM with AdaLN-Zero, reaching FC Corr 0.584, T. Corr 0.326, and MSE 0.239. This configuration improves connectivity while also exceeding MLP regression's pointwise accuracy.

Within these decoder configurations, the preferred velocity architecture also depends on the source and training objective. Under fixed-scale OT training, DiT reaches FC Corr 0.521 versus AdaLN-Zero's 0.442. With the learned source, β-NLL, and I-CFM held fixed, AdaLN-Zero reaches 0.584 versus DiT's 0.568. This reversal supports choosing the velocity architecture together with the source and training objective for connectivity recovery. Spectral-feature and electrode-coverage ablations are also reported in Appendix C.

Context window and seq2seq operating point. Beyond decoder design, the context sweep examines when predicting multiple BOLD volumes improves reconstruction. Earlier volumes in each predicted block have less preceding EEG context within the input window. At 16 s, four-volume prediction yields T. Corr 0.235 versus 0.281 for single-volume prediction; at 32 s, it reaches 0.326 versus 0.317 (Table 5, Appendix C). Extending context to 34 or 36 s gives no further gain. This supports the selected 32 s context and four-volume output, with overlapping blocks providing multiple estimates of each TR. The accompanying alignment analysis details the context available to each output.

Toward native uncertainty The distributional prior in Eq. (3) offers a path toward per-prediction uncertainty as a byproduct of the architecture (Table 2). Repeated predictions provide input-conditional ensemble spread without a separate uncertainty head, while connectivity retains its sampled-trajectory readout.

Diagnosis. On the fixed-scale pipeline, source noise, MC-Dropout, and deep ensembles yield weak error rankings and insufficient coverage. A heteroscedastic output head improves coverage but lowers T. Corr from 0.321 to 0.276. The transport diagnostic in Appendix E shows substantial contraction of source spread in the fixed-scale configuration; increasing its source scale only at inference preserves more spread but reduces temporal accuracy. These results motivate learning input-dependent source variation together with the flow.

Native predictive spread. The full configuration retains more source spread through transport and reaches Spearman 0.155, exceeding the fixed-scale alternatives in Table 2. The held-out analyses in Appendix E further characterize this signal: spread is higher in deep-gray and cerebellar components than in cortical components, and increases with temporal roughness of measured BOLD. At the connectivity level, selecting the least-uncertain half of FC edges reduces error by 6.9% $[2.0\%, 11.5\%]$ relative to a random half. This links ensemble spread to a practical use of the recovered connectomes: identifying connections with lower reconstruction error.

Calibration and complementary ranking. The raw ensemble provides useful error ordering, but its intervals remain too narrow. A single validation-fitted multiplier $\alpha{\approx}4.6$ raises nominal 95% coverage from 0.385 to $\mathbf{0.948}$ and reduces calibration error to $\mathbf{0.011}$, preserving the native ranking. Last-layer Laplace offers complementary ranking information: it reaches Spearman 0.195, and rank-average fusion with native spread reaches $\mathbf{0.225}$, above either alone and EP + Laplace. Scalar recalibration thus supports regional prediction intervals, while fusion improves error ordering using an additional uncertainty estimate.

5 Conclusion

BOLDFlow reframes EEG-to-fMRI synthesis as distributional, network-level prediction. It models $\hat{p}_\Omega(\mathbf{y} \mid \mathbf{x})$ over regional BOLD activity using conditional flow matching with a learned EEG-conditioned source. The covariance analysis connects this formulation to its connectivity readout: conditional-mean prediction omits residual cross-region covariance, while averaging sampled trajectories attenuates that component (Proposition 1 and Corollary 1). Joint generation and the sampled-trajectory readout retain the model's regional relationships for FC estimation.

Across subject-disjoint evaluations on NeuroBOLT and OpenNeuroSleep, BOLDFlow improves FC recovery together with temporal reconstruction. The gains extend to individual signatures and community organization, while input controls link recovery to the availability, identity, and temporal alignment of the supplied EEG. These results connect the distributional formulation to recovery of measured network structure across rest and sleep. Repeated predictions additionally provide native ensemble spread, supporting recalibrated regional prediction intervals and selection of FC edges with lower reconstruction error.

The evidence comes from simultaneous, in-scanner recordings of healthy adults. A central next step is to establish how well connectivity recovery and uncertainty calibration persist across larger cohorts, task states, and outside-scanner EEG. These extensions will determine how broadly the approach can support connectome-level analysis; Appendix I develops the limitations and broader implications.

AI USE STATEMENT

We used generative AI to assist with drafting and polishing the manuscript, searching for relevant literature, and providing feedback on mathematical arguments, and experimental controls during revision. The authors are responsible for the final text, analyses, and claims.

REPRODUCIBILITY STATEMENT

Architecture, optimization, prediction readouts and subject-disjoint evaluation are described in Appendices A–B; covariance derivations appear in Appendix G. Code: `https://anonymous.4open.science/r/boldflow-3FC6`.

REFERENCES

Sophie Achard and Edward T. Bullmore. Efficiency and cost of economical brain functional networks. *PLOS Computational Biology*, 3(2):e17, 2007. doi: 10.1371/journal.pcbi.0030017.

P. J. Allen, G. Polizzi, K. Krakow, D. R. Fish, and L. Lemieux. Identification of EEG events in the MR scanner: The problem of pulse artifact and a method for its subtraction. *NeuroImage*, 8(3): 229–239, 1998.

P. J. Allen, O. Josephs, and R. Turner. A method for removing imaging artifact from continuous EEG recorded during functional MRI. *NeuroImage*, 12(2):230–239, 2000.

Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

Vitalii Bondar, Serhii Semenov, Vira Babenko, and Dmytro Holovniak. FlowECG: Using flow matching to create a more efficient ECG signal generator. *arXiv preprint arXiv:2509.10491*, 2025.

Avishek Joey Bose, Tara Akhound-Sadegh, Guillaume Huguet, Kilian Fatras, Jarrid Rector-Brooks, Cheng-Hao Liu, Andrei Cristian Nica, Maksym Korablyov, Michael M. Bronstein, and Alexander Tong. SE(3)-stochastic flow matching for protein backbone generation. In *International Conference on Learning Representations*, 2024.

Lennart Bramlage and Cristóbal Curio. Principled input-output-conditioned post-hoc uncertainty estimation for regression networks. *arXiv preprint arXiv:2506.00918*, 2025.

David Calhas and Rui Henriques. EEG to fMRI synthesis benefits from attentional graphs of electrode relationships. In *Proceedings of the 8th Machine Learning for Healthcare Conference*, volume 219 of *Proceedings of Machine Learning Research*, pp. 76–93. PMLR, 2023.

Kamalaker Dadi, Gaël Varoquaux, Antonia Machlouzarides-Shalit, Krzysztof J. Gorgolewski, Demian Wassermann, Bertrand Thirion, and Arthur Mensch. Fine-grain atlases of functional modes for fMRI analysis. *NeuroImage*, 221:117126, 2020.

Erik Daxberger, Agustinus Kristiadi, Alexander Immer, Runa Eschenhagen, Matthias Bauer, and Philipp Hennig. Laplace redux—effortless Bayesian deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Maël Donoso. Neural networks and foundation models: Two strategies for EEG-to-fMRI prediction. *Frontiers in Systems Biology*, 5:1715692, 2025. doi: 10.3389/fsysb.2025.1715692.

Andrew T. Drysdale, Logan Grosenick, Jonathan Downar, Katharine Dunlop, Farrokh Mansouri, Yue Meng, Robert N. Fetcho, Benjamin Zebley, Desmond J. Oathes, Amit Etkin, Alan F. Schatzberg, Keith Sudheimer, Jennifer Keller, Helen S. Mayberg, Faith M. Gunning, George S. Alexopoulos, Michael D. Fox, Alvaro Pascual-Leone, Henning U. Voss, B. J. Casey, Marc J. Dubin, and Conor Liston. Resting-state connectivity biomarkers define neurophysiological subtypes of depression. *Nature Medicine*, 23(1):28–38, 2017.

Yassine El Ouahidi, Jonathan Lys, Philipp Thölke, Nicolas Farrugia, Bastien Pasdeloup, Vincent Gripon, Karim Jerbi, and Giulia Lioi. REVE: A foundation model for EEG – adapting to any setup with large-scale pretraining on 25,000 subjects. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 22541–22577. Curran Associates, Inc., 2025. doi: 10.52202/085713-0760.

Emily S. Finn, Xilin Shen, Dustin Scheinost, Monica D. Rosenberg, Jessica Huang, Marvin M. Chun, Xenophon Papademetris, and R. Todd Constable. Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nature Neuroscience*, 18(11):1664–1671, 2015.

Alex Fornito, Andrew Zalesky, and Michael Breakspear. The connectomics of brain disorders. *Nature Reviews Neuroscience*, 16(3):159–172, 2015.

Karl J. Friston. Functional and effective connectivity in neuroimaging: A synthesis. *Human Brain Mapping*, 2(1-2):56–78, 1994. doi: 10.1002/hbm.460020107.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning (ICML)*, 2016.

Gary H. Glover. Deconvolution of impulse response in event-related BOLD fMRI. *NeuroImage*, 9 (4):416–429, 1999.

Robin I. Goldman, John M. Stern, Jerome Engel Jr., and Mark S. Cohen. Simultaneous EEG and fMRI of the alpha rhythm. *NeuroReport*, 13(18):2487–2492, 2002.

Yameng Gu, Lucas E. Sainburg, Feng Han, and Xiao Liu. Simultaneous EEG and functional MRI data during rest and sleep from humans. *Data in Brief*, 48:109059, 2023.

Noam Issachar, Mohammad Salama, Raanan Fattal, and Sagie Benaim. Designing a conditional prior distribution for flow-based generative models. *Transactions on Machine Learning Research*, 2025.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. LaBraM: Large brain model for learning generic representations with tremendous EEG data in BCI. In *International Conference on Learning Representations (ICLR)*, 2024.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 5156–5165. PMLR, 2020.

Alex Kendall and Yarin Gal. What uncertainties do we need in Bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Marcel Kollovieh, Marten Lienen, David Lüdke, Leo Schwinn, and Stephan Günnemann. Flow matching with Gaussian process priors for probabilistic time series forecasting. In *International Conference on Learning Representations (ICLR)*, 2025.

Demetres Kostas, Stéphane Aroca-Ouellette, and Frank Rudzicz. BENDR: Using transformers and a contrastive self-supervised learning task to learn from massive amounts of EEG data. *Frontiers in Human Neuroscience*, 15:653659, 2021.

Alexander Kovalev, Ilia Mikheev, and Alexei Ossadtchi. fMRI from EEG is only deep learning away: the use of interpretable DL to unravel EEG-fMRI relationships. *arXiv preprint arXiv:2211.02024*, 2022.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2017.

Romeo Lanzino, Federico Fontana, Luigi Cinque, Francesco Scarcello, and Atsuto Maki. Neural transcoding vision transformers for EEG-to-fMRI synthesis. In *Computer Vision – ECCV 2024 Workshops*, volume 15642 of *Lecture Notes in Computer Science*, pp. 53–70. Springer, 2025. doi: 10.1007/978-3-031-91907-7_4.

Helmut Laufs, Andreas Kleinschmidt, Armin Beyerle, Evelyn Eger, Afraim Salek-Haddadi, Christine Preibisch, and Karsten Krakow. EEG-correlated fMRI of human alpha activity. *NeuroImage*, 19 (4):1463–1476, 2003.

Yamin Li, Ange Lou, Ziyuan Xu, Shengchao Zhang, Shiyu Wang, Dario J. Englot, Soheil Kolouri, Daniel Moyer, Roza G. Bayrak, and Catie Chang. Neurobolt: Resting-state EEG-to-fMRI synthesis with multi-dimensional feature mapping. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Xueqing Liu and Paul Sajda. A convolutional neural network for transcoding simultaneously acquired EEG-fMRI data. In *9th International IEEE/EMBS Conference on Neural Engineering*, pp. 477–482. IEEE, 2019.

Nikos K. Logothetis, Jon Pauls, Mark Augath, Trinath Trinath, and Axel Oeltermann. Neurophysiological investigation of the basis of the fMRI signal. *Nature*, 412(6843):150–157, 2001.

Dante Mantini, Mauro G. Perrucci, Cosimo Del Gratta, Gian Luca Romani, and Maurizio Corbetta. Electrophysiological signatures of resting state networks in the human brain. *Proceedings of the National Academy of Sciences*, 104(32):13170–13175, 2007.

Yehudit Meir-Hasson, Sivan Kinreich, Ilana Podlipsky, Talma Hendler, and Nathan Intrator. An EEG finger-print of fMRI deep regional activation. *NeuroImage*, 102:128–141, 2014.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *International Conference on Computer Vision (ICCV)*, 2023.

Ethan Perez, Florian Strub, Harm de Vries, Vincent Dumoulin, and Aaron C. Courville. FiLM: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Aram-Alexandre Pooladian, Heli Ben-Hamu, Carles Domingo-Enrich, Brandon Amos, Yaron Lipman, and Ricky T. Q. Chen. Multisample flow matching: Straightening flows with minibatch couplings. In *International Conference on Machine Learning (ICML)*, 2023.

Maximilian Seitzer, Arash Tavakoli, Dimitrije Antic, and Georg Martius. On the pitfalls of heteroscedastic uncertainty estimation with probabilistic neural networks. In *International Conference on Learning Representations (ICLR)*, 2022.

Stephen M. Smith, Thomas E. Nichols, Diego Vidaurre, Anderson M. Winkler, Timothy E. J. Behrens, Matthew F. Glasser, Kamil Ugurbil, Deanna M. Barch, David C. Van Essen, and Karla L. Miller. A positive-negative mode of population covariation links brain connectivity, demographics and behavior. *Nature Neuroscience*, 18(11):1565–1567, 2015.

Andrew Stirn, Harm Wessels, Megan Schertzer, Laura Pereira, Neville E. Sanjana, and David A. Knowles. Faithful heteroscedastic regression with neural networks. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

Matthew Tancik, Pratul P. Srinivasan, Ben Mildenhall, Sara Fridovich-Keil, Nithin Raghavan, Utkarsh Singhal, Ravi Ramamoorthi, Jonathan T. Barron, and Ren Ng. Fourier features let networks learn high frequency functions in low dimensional domains. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Alexander Tong, Kilian Fatras, Nikolay Malkin, Guillaume Huguet, Yanlei Zhang, Jarrid Rector-Brooks, Guy Wolf, and Yoshua Bengio. Improving and generalizing flow-based generative models with minibatch optimal transport. *Transactions on Machine Learning Research*, 2024.

Guangyu Wang, Wenchao Liu, Yuhong He, Cong Xu, Lin Ma, and Haifeng Li. EEGPT: Pretrained transformer for universal and reliable representation of EEG signals. In *Advances in Neural Information Processing Systems*, 2024.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *International Conference on Learning Representations*, 2025.

Cedric Huchuan Xia, Zongming Ma, Rastko Ciric, Shi Gu, Richard F. Betzel, Antonia N. Kaczkurkin, Monica E. Calkins, Philip A. Cook, Angel García de la Garza, Simon N. Vandekar, Zaixu Cui, Tyler M. Moore, David R. Roalf, Kosha Ruparel, Daniel H. Wolf, Christos Davatzikos, Ruben C. Gur, Raquel E. Gur, Russell T. Shinohara, Danielle S. Bassett, and Theodore D. Satterthwaite. Linked dimensions of psychopathology and connectivity in functional brain networks. *Nature Communications*, 9(1):3003, 2018.

Chaoqi Yang, M. Brandon Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In *Advances in Neural Information Processing Systems*, 2023.

Weiheng Yao, Zhihan Lyu, Mufti Mahmud, Ning Zhong, Baiying Lei, and Shuqiang Wang. CATD: Unified representation learning for EEG-to-fMRI cross-modal generation. *IEEE Transactions on Medical Imaging*, 44(7):2757–2767, 2025.

A Dataset Details and Model Development

NeuroBOLT. The resting-state dataset contains 22 healthy participants (mean age 35.2 ± 16.9 years, 12 female); seven contributed two scans, giving 29 scans (Li et al., 2024). Participants rested with their eyes closed throughout the scans. EEG was acquired at 5 kHz using an MR-compatible BrainAmp system with FCz reference; the analysis uses 26 scalp channels, excluding auxiliary EOG, EMG, and ECG channels. BOLD was acquired on a 3T Philips scanner with multi-echo EPI, TR$= 2.1$ s and TE$\in \{13, 31, 49\}$ ms. Dataset-specific details follow the original release.

OpenNeuroSleep. OpenNeuro ds003768 contains simultaneous EEG–fMRI recordings from 33 healthy participants (mean age 22.1 ± 3.2 years, 17 male/16 female) acquired during rest and sleep (Gu et al., 2023). Our evaluation uses 229 rest/sleep scans from this cohort. EEG was acquired at 5 kHz with FCz reference and AFz ground; we use 30 scalp channels. The fMRI acquisition used a 3T Siemens Prisma Fit with TR$= 2.1$ s, TE$= 25$ ms, 3×3 mm in-plane resolution, 4 mm slices, and 35 slices. The release includes sleep-stage annotations in 30-s epochs. The reported comparison pools the dataset's rest/sleep recordings.

Preprocessing. EEG undergoes gradient-artifact template subtraction (Allen et al., 2000) and ballistocardiogram correction (Allen et al., 1998), followed by 0.5 Hz high-pass filtering, resampling to 200 Hz, per-channel z-scoring, and clipping to ± 15 standard deviations. Context windows span 32 s and end at the corresponding fMRI trigger. fMRI preprocessing includes motion and slice-timing correction, spatial normalization, confound regression using motion parameters and their derivatives together with white-matter/CSF signals, and projection onto DiFuMo components (Dadi et al., 2020). After 0.15 Hz low-pass filtering, each BOLD component is demeaned within its scan and divided by the 95th percentile of its absolute centered values plus 10^{-8}, as implemented in the released preprocessing code. MSE is reported in these normalized target units. The trajectory audit compares outputs against the preprocessed targets, whose band limit is 0.15 Hz.

Subject-disjoint partitions and validation. Five fixed outer folds partition participants, keeping all scans from each participant together. Each NeuroBOLT fold holds out 4–6 test subjects and a separate set of three validation subjects, leaving 13–15 training subjects. Validation subjects determine early stopping and checkpoint selection; test subjects are reserved for evaluation. The five test partitions cover the 22 participants once. OpenNeuroSleep uses its own subject-disjoint partitions. Scalar uncertainty recalibration is fitted on held-out validation residuals. Three training seeds per outer fold yield 15 runs for the main comparison, with participant cohorts shared across seeds.

Development history and replication. Approximately 300 NeuroBOLT training runs preceded the final configuration: about 130 used the reported architecture, and about 170 explored earlier variants. These runs cover parcellations, ablations, seed repeats, and earlier configurations. Choices covered EEG context, output aggregation, velocity family/depth, source parameterization, and encoder configuration, using validation subjects within the fixed folds. In a five-fold checkpoint audit, selected epochs were 7/7/7/12/15 of the 30-epoch cap; for four folds retaining both checkpoints, later training worsened validation MSE by 0.049. These observations document the effect of validation-based checkpoint selection. OpenNeuroSleep was excluded from architecture, hyperparameter, and schedule selection. The configuration selected on NeuroBOLT was trained separately on OpenNeuroSleep, adjusting the input channel count. This second dataset therefore tests replication of a fixed configuration under a different acquisition distribution.

Data availability and ethics. NeuroBOLT is used under the terms of its original release; OpenNeuroSleep is available as ds003768. DiFuMo is used for academic research under its released license. The data are de-identified recordings collected with consent and ethics approval documented by the original providers. We collected no new human-subject data.

B Implementation Details

Architecture. The REVE temporal encoder is initialized from pretrained weights and fine-tuned end-to-end (El Ouahidi et al., 2025). It processes 200-sample EEG patches at stride 180, uses a

512-dimensional embedding, 22 transformer blocks, eight attention heads, and coordinate-based Fourier positional encoding. Attention pooling yields one temporal embedding. The spectral stream computes STFT features with window sizes $\{100, 200, 400, 800\}$ samples, projects and combines the scales, and uses four linear-attention layers with eight heads followed by mean pooling. The fused representation is $\mathrm{GELU}(\mathbf{h}_{\mathrm{temp}} + \mathbf{h}_{\mathrm{spec}})$.

The prior trunk has widths $512 \to 256 \to 128$ with GELU activations and separate mean and scale heads. Scales use $\mathrm{softplus}(\mathbf{s}) + 0.05$; the scale-head weights start at zero and its bias initializes the scale at 0.20. The velocity network uses width 512, four AdaLN-Zero blocks, and feed-forward width 2048. The default output is a flattened $4 \times R$ block; the sequence-to-one variant outputs R values. The reported model has 96.4M parameters, including 70.2M initialized from REVE. The total changes with the output dimension and ablation. Table 3 summarizes the training and inference configuration.

Table 3: **Training and inference configuration.**

Setting	Value
Optimizer / weight decay	AdamW / 0.01
Learning rate	10^{-4}, cosine decay to 10^{-6}
Warmup / epoch cap	3 epochs / 30 epochs
Checkpoint / early stopping	Validation Pearson correlation / patience 10
Batch size / gradient clipping	32 / max norm 2.0
Numerical precision	FP16 automatic mixed precision
Layer-wise learning-rate decay	None
Source regularizer	β-NLL, $\beta = 0.5$, $\lambda = 1$
Coupling	Matched-example I-CFM
EEG context / output block	32 s / 4 TRs
ODE solver	Euler, 50 steps
FC readout / UQ ensemble	$M = 1$ trajectory / $M = 50$ trajectories

Optimization. The REVE encoder and decoder use the same learning rate. Flow and prior losses jointly update the encoder, prior, and velocity parameters. Source samples are reparameterized without detaching their mean or scale from the flow objective. The auxiliary source loss and flow objective therefore jointly determine the learned source. The loss sums over target coordinates in its mathematical definition; a consistent coordinate average can equivalently be used for both terms in implementation.

Inference and aggregation. The default FC prediction draws a Gaussian source for every anchor, transports it through the flow, and overlap-averages the predicted blocks. One full set of such draws is one sampled trajectory ($M = 1$). Repeating the full procedure independently M times provides both an ensemble-mean readout and an uncertainty estimate. The ensemble-mean readout averages per-TR predictions before computing temporal or FC metrics. UQ instead uses their mean and Bessel-corrected standard deviation at each output coordinate to construct regional intervals, while the FC readout remains a sampled trajectory.

A separate deterministic readout sets $\mathbf{z}_0 = \boldsymbol{\mu}_\phi(\mathbf{c})$ before integration. Runtime measurements in the original implementation used this deterministic readout: 127.7 ± 9.4 ms per window at batch size 1, or 16.5 ms per window at batch size 32, on an A100-PCIE-40GB with 50 Euler steps. A 50-member ensemble at batch size 1 took 4.68 ± 0.18 s, with the encoder evaluated once per window (200 timed trials after 20 warmup trials).

Compute. Each run uses one A100 40 GB GPU and takes approximately 2.5 hours. Training one configuration across five folds and three seeds therefore takes approximately 37.5 GPU-hours; the broader development effort is described in Appendix A.

Baseline definitions. The comparisons separate the per-region NeuroBOLT formulation (Li et al., 2024), a joint-output adaptation, and a NeuroBOLT-encoder/BOLDFlow-decoder hybrid. REVE-NoFT and REVE-FT use an MLP readout with frozen and fine-tuned REVE, respectively. The

Algorithm 1 BOLDFlow training and trajectory readouts.

Require: Matched EEG windows and BOLD blocks; $\beta = 0.5$, $\lambda = 1$.
Training: encode $\mathbf{c} = \mathrm{Enc}_\psi(\mathbf{x})$ and obtain $(\boldsymbol{\mu}, \boldsymbol{\sigma}) = \mathrm{Prior}_\phi(\mathbf{c})$.
Draw $\boldsymbol{\epsilon} \sim \mathcal{N}(0, I)$ and $\tau \sim \mathcal{U}[0, 1]$; set $\mathbf{z}_0 = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$.
Form $\mathbf{z}_\tau = (1 - \tau)\mathbf{z}_0 + \tau\mathbf{y}$ and velocity target $\mathbf{y} - \mathbf{z}_0$.
Update all parameters using Eqs. (4) and (5).
Inference: encode each scan's EEG windows; choose repetition count M.
for $m = 1, \ldots, M$ **do**
  for each anchor t **do**
    Draw a fresh source $\mathbf{z}_0^{(m)}$ from the anchor's learned Gaussian.
    Integrate 50 Euler steps and reshape to a $T_{\mathrm{out}} \times R$ block.
  end for
  Overlap-average blocks into trajectory $\hat{\mathbf{Y}}^{(m)}$.
end for
Sampled-trajectory FC: compute FC from $\hat{\mathbf{Y}}^{(1)}$.
Ensemble-mean FC: compute FC after averaging the M trajectories.
UQ: compute per-TR, per-component ensemble mean and standard deviation; apply validation-fitted scalar recalibration to the latter.

BOLDFlow† row in the parcellation table denotes the BOLDFlow encoder pipeline with a joint MSE-trained MLP head; it is distinct from NeuroBOLT†. Main-comparison methods use the same subject partitions and evaluation metrics. Context and output-aggregation effects are assessed separately in Table 5; these sweeps complement the encoder and decoder comparisons.

C Architecture and Temporal Context

Positional encoding. REVE's coordinate-based encoding represents electrode position and patch time. We train separately on each dataset, using this encoding to represent its input montage.

Velocity blocks. For hidden state $\mathbf{h}$ and a modulation vector $\mathbf{m}$ obtained from the EEG embedding and sinusoidal flow-time embedding, each block computes

$$(\boldsymbol{\gamma}, \boldsymbol{\beta}, \boldsymbol{a}) = \mathrm{MLP}_{\mathrm{mod}}(\mathbf{m}), \quad \tilde{\mathbf{h}} = (1 + \boldsymbol{\gamma}) \odot \mathrm{LN}(\mathbf{h}) + \boldsymbol{\beta}, \quad \mathbf{h} \leftarrow \mathbf{h} + \boldsymbol{a} \odot \mathrm{FFN}(\tilde{\mathbf{h}}). \tag{7}$$

Zero-initialized modulation makes the residual blocks initially act as identities. A final projection returns the predicted velocity in the target block space.

Alternative fusion. The gated cross-attention ablation uses learnable queries over temporal and spectral representations, followed by a sigmoid gate and a feed-forward layer. Additive fusion yields higher FC Corr at all three tested parcellations (Table 4), and remains the default. The small MSE advantage of gated fusion illustrates that the preferred configuration depends on the evaluation target.

Table 4: **Fusion strategy across parcellation granularities** on NeuroBOLT dataset with **BOLDFlow**. Best per parcellation in **bold**.

Parcellation	Fusion	T. Corr. ↑	MSE ↓	FC Corr. ↑
DiFuMo-64	Additive	**0.326**$_{\pm 0.008}$	0.239$_{\pm 0.002}$	**0.584**$_{\pm 0.017}$
	Gated cross-attention	0.307$_{\pm 0.009}$	**0.236**$_{\pm 0.003}$	0.492$_{\pm 0.014}$
DiFuMo-256	Additive	**0.288**$_{\pm 0.012}$	0.252$_{\pm 0.002}$	**0.620**$_{\pm 0.018}$
	Gated cross-attention	0.278$_{\pm 0.008}$	**0.244**$_{\pm 0.004}$	0.528$_{\pm 0.015}$
DiFuMo-512	Additive	0.256$_{\pm 0.010}$	0.262$_{\pm 0.003}$	**0.550**$_{\pm 0.020}$
	Gated cross-attention	**0.267**$_{\pm 0.009}$	**0.251**$_{\pm 0.002}$	0.486$_{\pm 0.018}$

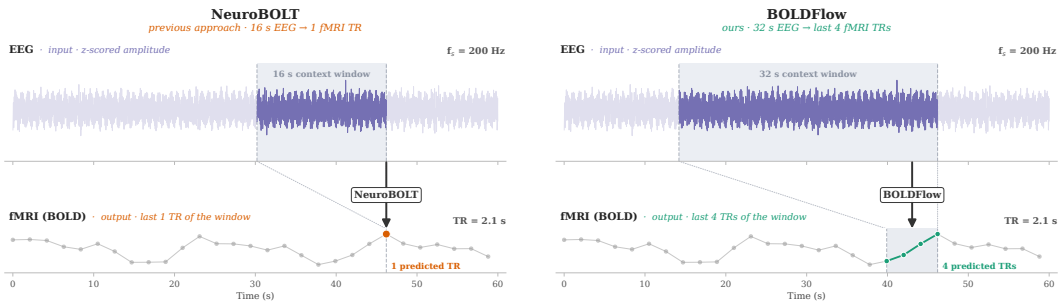

Figure 3: **Temporal context and target blocks.** NeuroBOLT's original sequence-to-one configuration uses 16 s of EEG. BOLDFlow predicts four trailing TRs from 32 s; overlapping blocks are averaged during offline reconstruction.

Context and overlap. For a target TR t, the $K_t \leq T_{\text{out}}$ blocks covering it contribute

$$\hat{\mathbf{y}}_t^{\text{agg}} = \frac{1}{K_t} \sum_{k:\, t \in \text{block}(k)} \hat{\mathbf{y}}_t^{(k)}. \tag{8}$$

Conditioned on the contributing EEG windows, independent anchor draws contribute no cross-anchor covariance, so the conditional covariance of the average is the sum of their conditional covariances divided by K_t^2. Across EEG inputs, the predicted means can covary because their contexts overlap. Outputs at different TRs can also share a sampled block. Appendix G derives the resulting covariance decomposition used in Corollary 1.

The earliest output in a block ending at its EEG anchor has only

$$\Delta_{\text{eff}} = \Delta_{\text{ctx}} - (T_{\text{out}} - 1)\text{TR}$$

seconds of preceding EEG in that window. With 32 s context and four outputs, this is 25.7 s. Figure 3 illustrates the alignment of EEG context and target blocks. The complete overlap-aggregated estimate of a TR can also use EEG up to three TRs after that TR; the reported pipeline performs offline reconstruction.

Table 5 reports temporal accuracy and between-run dispersion across context lengths. At 16 s, sequence-to-sequence underperforms sequence-to-one on T. Corr (0.235 versus 0.281); at 32 s, the ordering reverses (0.326 versus 0.317). Sequence-to-one peaks at 28 s (0.320 ± 0.017) and sequence-to-sequence at 32 s (0.326 ± 0.008). Extending the window to 34 or 36 s gives no further gain in either formulation. The sweep supports the 32-s sequence-to-sequence operating point and shows that its benefit depends on the available EEG context.

Table 5: **Context-window sweep, seq2one vs. seq2seq.** Temporal Pearson correlation as a function of the EEG input length Δ_{ctx}, at step 2 s from 16 to 36 s. All runs on NEUROBOLT at DiFuMo-64. The seq2seq columns use T_{out}=4 and the overlapping-window aggregation in Eq. (8), with $K_t \leq T_{\text{out}}$ contributing blocks at TR t.

Window Δ_{ctx}	**Seq2one** $r\uparrow$	**Seq2seq** T_{out}=4 $r\uparrow$
16 s	$0.281_{\pm 0.013}$	$0.235_{\pm 0.011}$
18 s	$0.291_{\pm 0.014}$	$0.265_{\pm 0.012}$
20 s	$0.310_{\pm 0.012}$	$0.285_{\pm 0.012}$
22 s	$0.306_{\pm 0.015}$	$0.303_{\pm 0.011}$
24 s	$0.302_{\pm 0.017}$	$0.310_{\pm 0.011}$
26 s	$0.311_{\pm 0.017}$	$0.318_{\pm 0.010}$
28 s	$\mathbf{0.320}_{\pm 0.017}$	$0.322_{\pm 0.009}$
30 s	$0.318_{\pm 0.015}$	$0.324_{\pm 0.009}$
32 s	$0.317_{\pm 0.013}$	$\mathbf{0.326}_{\pm 0.008}$
34 s	$0.316_{\pm 0.013}$	$0.325_{\pm 0.008}$
36 s	$0.312_{\pm 0.014}$	$0.322_{\pm 0.008}$

Table 6: **Decoder ablations on NeuroBOLT.** The encoder is REVE with the spectral branch, additive fusion, attention pooling, and 32-s context. L0/L0′ compare linear/MLP regression. L1–L3 change the source, coupling, and auxiliary loss under MLP velocity. The middle block changes velocity architecture at the L3 configuration; the bottom block compares velocity networks under learned-source I-CFM with β-NLL. The full-versus-L4 comparison changes the source/loss/coupling bundle, rather than one factor.

Configuration	T. Corr. ↑	MSE ↓	FC Corr. ↑	Δ FC vs L0
Decoder family (MLP velocity):				
L0: Linear regression	$0.299_{\pm 0.013}$	$0.243_{\pm 0.002}$	$0.397_{\pm 0.018}$	–
L0′: MLP regression	$0.315_{\pm 0.012}$	$0.241_{\pm 0.002}$	$0.420_{\pm 0.017}$	+0.023
L1: Gaussian source, I-CFM	$0.281_{\pm 0.013}$	$0.252_{\pm 0.002}$	$0.461_{\pm 0.016}$	+0.064
L2: Gaussian source, OT	$0.293_{\pm 0.012}$	$0.249_{\pm 0.002}$	$0.470_{\pm 0.016}$	+0.073
L3: Fixed-σ source, OT	$0.299_{\pm 0.012}$	$0.246_{\pm 0.002}$	$0.452_{\pm 0.015}$	+0.055
Velocity architecture (source, coupling, and loss fixed at L3):				
L3 + DiT	$0.304_{\pm 0.010}$	$0.243_{\pm 0.002}$	$0.521_{\pm 0.015}$	+0.124
L4: L3 + AdaLN-Zero	$0.319_{\pm 0.012}$	$\mathbf{0.234}_{\pm 0.002}$	$0.442_{\pm 0.018}$	+0.045
Learned source, I-CFM, and β-NLL:				
DiT + learned source	$0.318_{\pm 0.011}$	$0.241_{\pm 0.002}$	$0.568_{\pm 0.017}$	+0.171
Full BOLDFlow (AdaLN-Zero)	$\mathbf{0.326}_{\pm 0.008}$	$0.239_{\pm 0.002}$	$\mathbf{0.584}_{\pm 0.017}$	**+0.187**

Decoder ablations. *Decoder family.* With REVE and the spectral branch held fixed, Gaussian-source CFM improves FC Corr over MLP regression, from 0.420 to 0.461 with I-CFM or 0.470 with OT coupling, despite lower T. Corr and higher MSE (Table 6). These alternatives demonstrate a trade-off between connectivity and pointwise accuracy within a shared EEG representation. Full BOLDFlow reaches FC Corr 0.584 while also improving T. Corr and MSE over MLP regression, combining gains in network recovery and regional reconstruction.

Source and velocity design. The preferred velocity architecture depends on the source and training configuration. Under fixed-scale OT training, DiT attains higher FC Corr than AdaLN-Zero (0.521 versus 0.442), while AdaLN-Zero achieves the lowest MSE in Table 6. Combining input-dependent source scale, auxiliary supervision, and I-CFM improves FC under both architectures and reverses their ordering: AdaLN-Zero reaches 0.584 and DiT 0.568. This reversal supports selecting the source, training objective, and velocity architecture jointly for the connectivity endpoint.

Spectral branch and montage. The final-stack spectral-branch ablation reaches FC Corr 0.547 without B2, compared with 0.584 with B2; removing B2 also increases MSE by 0.009. This isolates the contribution of the NeuroBOLT-style spectral stream within the full decoder configuration.

Reduced-montage models were retrained with correct electrode coordinates under the same five-fold protocol, retaining both encoder streams. The 19-channel configuration uses the standard 10–20 subset, and the six-channel configuration uses F3/F4/C3/C4/O1/O2. These six channels retain the original reference. Table 7 reports temporal and connectivity recovery for each montage. The

Table 7: **Reduced-montage results.** Point estimates under the five-fold subject-disjoint protocol.

Channels	T. Corr	FC Corr
26	0.326	0.584
19	0.305	0.518
6	0.241	0.463

19-channel temporal value follows the reported 0.021 drop from the 26-channel configuration. Both metrics decrease as the input montage is reduced.

Output parcellation. BOLDFlow has the highest FC Corr among the compared methods at 64, 256, and 512 components (Table 8), reaching 0.550 versus NeuroBOLT's 0.433 at 512 components. Finer parcellations reduce temporal correlation for all compared methods, while BOLDFlow maintains

a T. Corr advantage of approximately 0.05 over NeuroBOLT at each granularity. Thus, its relative benefit persists as the output representation becomes finer. These targets remain strongly correlated: at 512 components, measured-fMRI effective rank is 7.9 [6.3, 9.7], or 7.2 [5.7, 8.7] after excluding 90 predominantly non-neural components. A 512-specific pairing test is inconclusive; the corresponding DiFuMo-64 control is reported in Appendix D.

Table 8: **Parcellation scaling.** Per-method, per-granularity results on NEUROBOLT dataset. The two rightmost columns report the relative change from DiFuMo-64 to DiFuMo-256 and from DiFuMo-64 to DiFuMo-512: these are descriptive changes rather than a separate ranking criterion. BOLDFlow† shares the BOLDFlow encoder pipeline but swaps the CFM decoder for an MSE-trained MLP head, corresponding to L0$'$ in Table 6 at 64 components. Best per metric column in **bold**.

Method	Metric	DiFuMo-64	DiFuMo-256	DiFuMo-512	Δ_{rel} (64→256)	Δ_{rel} (64→512)
BOLDFlow (ours)	T. Corr. ↑	**0.326**$_{\pm 0.008}$	**0.288**$_{\pm 0.012}$	**0.256**$_{\pm 0.010}$	−12%	−21%
	MSE ↓	**0.239**$_{\pm 0.002}$	**0.252**$_{\pm 0.002}$	**0.262**$_{\pm 0.003}$	+5%	+10%
	FC Corr. ↑	**0.584**$_{\pm 0.017}$	**0.620**$_{\pm 0.018}$	**0.550**$_{\pm 0.020}$	+6%	−6%
BOLDFlow†	T. Corr. ↑	0.315$_{\pm 0.012}$	0.265$_{\pm 0.012}$	0.228$_{\pm 0.011}$	−16%	−28%
	MSE ↓	0.241$_{\pm 0.002}$	0.254$_{\pm 0.002}$	0.264$_{\pm 0.003}$	+5%	+10%
	FC Corr. ↑	0.420$_{\pm 0.017}$	0.435$_{\pm 0.019}$	0.424$_{\pm 0.022}$	+4%	+1%
NeuroBOLT	T. Corr. ↑	0.272$_{\pm 0.013}$	0.234$_{\pm 0.012}$	0.205$_{\pm 0.010}$	−14%	−25%
	MSE ↓	0.246$_{\pm 0.002}$	0.255$_{\pm 0.002}$	0.265$_{\pm 0.003}$	+4%	+8%
	FC Corr. ↑	0.428$_{\pm 0.014}$	0.427$_{\pm 0.018}$	0.433$_{\pm 0.020}$	0%	+1%

D INPUT SPECIFICITY AND TRAJECTORY DIAGNOSTICS

The input controls evaluate held-out NeuroBOLT recordings under five subject-disjoint folds, covering 22 subjects and 29 scans. They use the sampled-trajectory readout, with the additional ensemble-mean comparison specified below.

Constant-input generator. The same architecture is retrained with EEG replaced by zero under the subject-disjoint protocol. Its sampled-trajectory FC Corr is 0.264 versus 0.584 for BOLDFlow. The constant-input model's FC remains similar after trajectory averaging, with a large-ensemble value of 0.255, whereas BOLDFlow reaches 0.471 at $M = 200$ in Table 14. A time-invariant conditional mean removes the across-input covariance term, so the population correlation matrix of a residual ensemble with positive component variances is invariant to uniform covariance scaling by $1/M$ for finite M. Finite-scan estimates also reflect sampling variability and overlap boundary effects.

Pairing beyond population structure. The training-population FC template is removed before comparing correct- and wrong-subject EEG. Correct-pair residual FC correlation is 0.146; the reported 95% interval for the matched advantage is [0.070, 0.280], and its sign is positive in all five folds. A matched-minus-wrong interaction with the constant-input generator is also positive in the non-residual analysis (interval [0.007, 0.163]); the population-residual observed-minus-null interaction is inconclusive. The pairing contrast evaluates recovery of scan-specific connectivity after removing the shared template, complementing the constant-input comparison and the within-scan alignment test.

Within-scan temporal specificity. Dynamic FC is computed in 120-s windows. Aligned EEG outperforms EEG displaced by 180 s, with a reported interval for the contrast of [0.016, 0.041]. The correct-versus-different-subject dynamic contrast has interval [0.058, 0.172]. Subjects are the resampling unit for these temporal-window statistics. The contrasts quantify sensitivity to subject pairing and alignment within the observed recordings.

Full held-out trajectory audit. We drew 200 trajectories per held-out EEG input across five folds (29 scans, 22 subjects), evaluated individual generated samples, and bootstrapped subjects. Table 9 distinguishes nonzero conditional diversity from matching the target's temporal distribution. Effective rank summarizes how covariance is distributed across eigenmodes.

Table 9: **Diversity and calibration diagnostics.** Brackets are reported subject-bootstrap 95% intervals. Sample/target variance and FC-effective-rank ratios equal one when the corresponding statistics match.

Diagnostic	Result
Terminal/source RMS spread	0.383 [0.362, 0.405]
Prior scales at numerical floor	0.0%
Pairwise sample RMS / target SD	0.340 [0.323, 0.360]
Conditional-covariance effective rank	31.6/64
Conditional effective rank / dimension	0.494 [0.472, 0.517]
Sample/target temporal variance	0.274 [0.217, 0.343]
Sample/target across-component variance	0.177 [0.148, 0.212]
Sample/target FC effective rank	0.643 [0.520, 0.805]
In-band spectral total-variation distance	0.095 [0.077, 0.116]
Lag-1 autocorrelation: generated / measured	0.557 / 0.810
Power above 0.15 Hz: generated / measured	13.0% / 0.1%

The outputs have multidimensional conditional spread, compressed amplitudes and network spectra, and lower temporal autocorrelation than the targets. Pearson FC is invariant to a constant positive rescaling of each component, allowing amplitude and correlation structure to be assessed separately.

Filtering control. Generated series were low-pass filtered at 0.15 Hz, the band limit already present in the measured targets, before recomputing FC across the same held-out scans. The reported changes in both the single-draw and ensemble analyses have intervals containing zero and exclude a loss greater than 0.014 despite removal of 13.0% of generated power. This quantifies the stability of the FC result under matching of the target frequency band. A 0.01–0.15 Hz bandpass was also examined, but additionally removes slow content from both series and addresses a different target statistic.

E Uncertainty and Connectivity Quality Control

Output spread and scalar recalibration. For $M = 50$ independent trajectory realizations, define $\hat{\mu}_{t,r} = M^{-1} \sum_m \hat{y}_{t,r}^{(m)}$ and

$$s_{t,r} = \left(\frac{1}{M-1} \sum_m (\hat{y}_{t,r}^{(m)} - \hat{\mu}_{t,r})^2 \right)^{1/2}.$$

For native ensemble UQ, the prediction center is $\hat{\mu}_{t,r}$ and the uncertainty readout is $\alpha s_{t,r}$, with α fitted on validation residuals. At nominal level q, a Gaussian marginal interval is

$$I_{t,r}(q) = \left[\hat{\mu}_{t,r} - z_{(1+q)/2} \alpha s_{t,r},\ \hat{\mu}_{t,r} + z_{(1+q)/2} \alpha s_{t,r} \right],$$

where z_p is the standard-normal quantile. Coverage is the fraction of evaluated targets inside the corresponding interval; Spearman correlation ranks absolute residuals against uncertainty scores. The raw and scalar-recalibrated ensemble use the same prediction center, so positive scaling changes coverage but not rankings.

Table 2 reports error ranking, calibration error, and nominal 95% interval coverage. Coverage evaluates interval width against observed residuals, while error ranking evaluates the ordering of predictions by uncertainty.

The raw ensemble covers 38.5% of targets at the nominal 95% level, increasing to 94.8% after scalar recalibration. Combining the ranks supplied by native spread and last-layer Laplace raises the ranking score to 0.225, compared with 0.155 and 0.195 for the individual estimates.

Edge-error selection. The held-out ensemble-spread analysis ranks errors at FC-edge, component, and time-point levels, with positive associations in each fold and Holm-significant reported tests. Selecting the least-uncertain half of FC edges reduces error by 6.9% [2.0%, 11.5%] relative to selecting a random half. The retention fraction is fixed at one half for this comparison.

Fixed-scale comparisons and ensemble diversity. On the fixed-σ pipeline, source noise, MC-Dropout, and deep ensembles provide weak error ranking, with Spearman correlations between 0.005 and 0.045; the heteroscedastic head lowers T. Corr from 0.321 to 0.276 (Table 2). Table 10 examines how source variation survives transport. Across 512 validation windows, the learned-source configuration retains 29% of source spread, compared with 7% for the fixed-scale configuration at its training scale, with terminal standard deviations of 0.092 and 0.007, respectively. Increasing the fixed source scale only at inference retains more spread but lowers T. Corr. This validation-window diagnostic measures mean coordinatewise standard deviation; the held-out trajectory audit in Table 9 separately reports RMS spread over full recordings. The diagnostics characterize retained variation, while Table 2 evaluates error ranking and calibration.

Table 10: **Source-spread preservation through the ODE flow on NEUROBOLT** (DiFuMo-64, $M{=}20$ source samples per window, ensemble std averaged over 64 ROIs and 512 validation windows). *Fixed-σ prior*: the source is fixed as $\mathcal{N}(\boldsymbol{\mu}_{\psi,\phi}, \sigma_{\mathrm{src}}^2\mathbf{I})$ with σ_{src} a hyperparameter; $\sigma_{\mathrm{src}}{=}0.1$ is the training regime, larger values are tested only at inference. *Distributional prior*: BOLDFlow's learned per-sample $\boldsymbol{\sigma}_\phi(\mathbf{x})$ supervised by β-NLL replaces the hyperparameter schedule. *Spread preserved* $:= \mathrm{std}(\tau{=}1)/\mathrm{std}(\tau{=}0)$. Off-regime σ_{src} values preserve more source spread, but the velocity field was never trained at those regimes and per-TR T. Corr drops accordingly.

Prior configuration	**std**($\tau{=}0$)	**std**($\tau{=}1$)	**Spread preserved**	**T. Corr**
Fixed-σ, $\sigma_{\mathrm{src}}{=}0.1$ (training regime)	0.100	$0.007_{\pm 0.001}$	$0.07_{\pm 0.01}$	$0.321_{\pm 0.012}$
Fixed-σ, $\sigma_{\mathrm{src}}{=}0.3$ (off-regime)	0.300	$0.040_{\pm 0.003}$	$0.13_{\pm 0.01}$	$0.298_{\pm 0.013}$
Fixed-σ, $\sigma_{\mathrm{src}}{=}0.5$ (off-regime)	0.500	$0.105_{\pm 0.008}$	$0.21_{\pm 0.02}$	$0.262_{\pm 0.013}$
Fixed-σ, $\sigma_{\mathrm{src}}{=}1.0$ (off-regime)	1.000	$0.305_{\pm 0.020}$	$0.30_{\pm 0.02}$	$0.187_{\pm 0.014}$
Distributional prior (learned $\boldsymbol{\sigma}_\phi$)	$0.314_{\pm 0.018}$	$0.092_{\pm 0.007}$	$0.29_{\pm 0.02}$	$0.326_{\pm 0.011}$

Regional and temporal associations. This anatomical audit uses a 56-cortical/5-deep-gray grouping, distinct from the fixed 55-component cortical mask used for the primary FC metric. Across these 61 neural DiFuMo-64 components, spread is higher in the five deep-gray/cerebellar components than in the 56 cortical components (0.138 versus 0.117). Spread also increases with temporal roughness of measured BOLD; rank-partialling target standard deviation strengthens the reported associations. The association with 0.08–0.15 Hz power is not significant under the permutation test. These associations characterize regional differences in predictive spread under paired EEG–BOLD supervision.

F NETWORK ANALYSES AND CROSS-SESSION PERSONALIZATION

Group-level FC visualization. Per-scan FC matrices are transformed entrywise to Fisher-z, averaged across 29 held-out scans, and transformed back. Edge density is the fraction of possible undirected connections retained in a thresholded graph (Achard & Bullmore, 2007). Figure 1 displays the 100 largest absolute measured-FC edges, approximately 5% of the 2016 possible DiFuMo-64 edges. An edge is recovered if it also appears in the prediction's top 100. This descriptive figure uses the full 64-component matrix; the headline FC Corr uses the fixed 55-component cortical evaluation mask. Cortical network labels organize the display. Per-network similarities summarize edges incident to each network and describe group-level structure.

Community recovery. Predicted and measured FC matrices are thresholded at matched 5% density. Louvain community detection uses resolution $\gamma = 1$, and adjusted Rand index compares the resulting partitions, averaged over held-out scans (Table 11). BOLDFlow achieves ARI 0.301, compared with 0.118 for joint-MSE and 0.068 for per-ROI NeuroBOLT. With the number of retained connections held fixed, the higher ARI indicates that their placement organizes regions into communities closer to those in measured fMRI. This complements the edge-recovery analysis in Figure 1 by evaluating how the recovered connections partition the network.

Identification within predicted connectomes. Each scan is split into two non-overlapping halves. First-half predicted FC is matched against second-half predicted FC by upper-triangle Pearson correlation, and the direction is reversed for the symmetrized score. The pool contains 22 participants;

Table 11: **Louvain community-structure preservation on predicted FC** (NEUROBOLT, DiFuMo-64). Both predicted and ground-truth FC matrices are thresholded at matched edge density (top 5%); Louvain community detection is run on each with resolution γ=1; the adjusted Rand index (ARI) measures partition agreement, averaged across held-out scans. ARI $= 0$ corresponds to partition agreement at chance level by construction.

FC source	ARI (predicted vs. true) ↑
BOLDFlow	$0.301_{\pm 0.025}$
NeuroBOLT	$0.118_{\pm 0.015}$
NeuroBOLT†	$0.068_{\pm 0.011}$

chance is $1/22$. Table 12 compares identification accuracy from predicted scan halves with the measured-to-measured reference. BOLDFlow reaches 84.1%, versus 72.7% for joint-MSE and 70.5% for per-ROI NeuroBOLT, indicating more stable individual network signatures across predicted scan halves.

Table 12: **Subject identification accuracy from predicted functional-connectivity matrices** (NEUROBOLT, N=22 held-out subjects, chance $= 1/22 \approx 4.5\%$). For each subject, the fMRI scan is split into two non-overlapping halves; FC is computed on each half (predicted-to-predicted within each model, measured-to-measured for the reference); top-1 identification matches each subject's first half against the pool of all subjects' second halves via Pearson correlation of the upper-triangular FC entries (symmetrised over query/target directions).

Source of FC matrices	Top-1 accuracy ↑	Chance-adjusted
Measured FC halves (reference)	90.9%	×20.0
BOLDFlow predicted FC	84.1%	×18.5
NeuroBOLT predicted FC	72.7%	×16.0
NeuroBOLT† per-ROI predicted FC	70.5%	×15.5

Cross-session personalization. The seven two-scan subjects are evaluated in both adaptation/test directions (14 directions). Each starting checkpoint comes from the outer fold excluding that subject, so neither scan entered its original training. We freeze the EEG feature path (REVE, spectral stream, and fusion), adapt the velocity decoder and, when selected, its prior head on one scan, and evaluate on the other. A held-out tail of the adaptation scan, separated by a 10-frame gap, controls early stopping. Learning rate and adaptation scope are selected by leave-one-subject-out cross-fitting over the other six participants.

Zero-shot FC Corr on these subjects is 0.570. Personalization improves FC in all 14 directions, with a subject-clustered interval for the improvement of [0.077, 0.293]. Temporal correlation changes from 0.322 to 0.331 and MSE from 0.239 to 0.237, without a clear change under the reported uncertainty intervals. The improvement is evaluated on the independent test scan after adaptation on the subject's other recording.

Regional accuracy and the thalamus. Table 13 reports per-component temporal correlations, with BOLDFlow improving over per-ROI NeuroBOLT in 52 of the 56 cortical components and two of the five deep-gray/cerebellar components. The regional pattern favors BOLDFlow across most cortical targets, while per-ROI NeuroBOLT retains an advantage in the thalamus and both cerebellar components. The per-region NeuroBOLT thalamic correlation is 0.317 versus 0.216 for BOLDFlow. A joint-output NeuroBOLT control reaches 0.210, close to BOLDFlow. This supports a role for per-region specialization in the relative gap. Measured thalamic lag-1 autocorrelation is 0.763 versus 0.818 for cortex, a post-hoc association.

Representative time courses. Figure 4 shows nine components in a held-out subject. The examples illustrate variation in temporal agreement across components, complementing the multi-sample audit in Table 9.

Table 13: **Per-ROI prediction quality** on NEUROBOLT (DiFuMo-64, inter-subject 5-fold CV). BOLDFlow predicts all 64 ROIs jointly; NeuroBOLT† trains one model per ROI. Temporal correlation (Pearson r) averaged across folds. $\Delta > 0$: BOLDFlow outperforms. Best per ROI in **bold**. *Non-neural component.

ROI	BOLDFlow ↑	NeuroBOLT† ↑	Δ	Network
Superior parts of Postcentral and Precentral gyri	**0.434**$_{\pm 0.014}$	0.365$_{\pm 0.015}$	+0.069	SomMot
Inferior occipital gyrus	**0.432**$_{\pm 0.012}$	0.346$_{\pm 0.017}$	+0.086	Vis
Calcarine cortex posterior	**0.423**$_{\pm 0.013}$	0.355$_{\pm 0.012}$	+0.068	Vis
Callosomarginal sulcus	**0.420**$_{\pm 0.013}$	0.376$_{\pm 0.014}$	+0.044	SalVAttn
Superior parietal lobule anterior	**0.419**$_{\pm 0.014}$	0.355$_{\pm 0.017}$	+0.064	DorsAttn
Paracentral gyrus RH	**0.414**$_{\pm 0.015}$	0.353$_{\pm 0.015}$	+0.061	SomMot
Precuneus superior	**0.413**$_{\pm 0.016}$	0.343$_{\pm 0.016}$	+0.070	DorsAttn
Paracentral lobule superior	**0.412**$_{\pm 0.019}$	0.343$_{\pm 0.017}$	+0.069	SomMot
Superior occipital gyrus	**0.412**$_{\pm 0.008}$	0.337$_{\pm 0.011}$	+0.075	Vis
Cuneus	**0.411**$_{\pm 0.011}$	0.339$_{\pm 0.009}$	+0.072	Vis
Calcarine sulcus anterior	**0.410**$_{\pm 0.009}$	0.353$_{\pm 0.009}$	+0.057	Vis
Heschl's gyrus	0.409$_{\pm 0.017}$	**0.410**$_{\pm 0.012}$	−0.001	SomMot
Central sulcus	**0.408**$_{\pm 0.015}$	0.346$_{\pm 0.012}$	+0.062	SomMot
Lingual gyrus	**0.407**$_{\pm 0.011}$	0.334$_{\pm 0.014}$	+0.073	Vis
Lingual gyrus anterior	**0.403**$_{\pm 0.011}$	0.368$_{\pm 0.008}$	+0.035	Vis
Parieto-occipital sulcus middle	**0.398**$_{\pm 0.011}$	0.343$_{\pm 0.009}$	+0.055	Vis
Postcentral gyrus LH	**0.395**$_{\pm 0.016}$	0.327$_{\pm 0.014}$	+0.068	SomMot
Planum temporale	**0.393**$_{\pm 0.016}$	0.379$_{\pm 0.012}$	+0.014	SomMot
Descending occipital gyrus	**0.389**$_{\pm 0.009}$	0.312$_{\pm 0.012}$	+0.077	Vis
Fusiform gyrus posterior	**0.383**$_{\pm 0.013}$	0.275$_{\pm 0.010}$	+0.108	Vis
Superior parietal lobule	**0.382**$_{\pm 0.013}$	0.308$_{\pm 0.018}$	+0.074	DorsAttn
Precuneus anterior	**0.379**$_{\pm 0.015}$	0.332$_{\pm 0.018}$	+0.047	DorsAttn
Lateral occipital cortex	**0.378**$_{\pm 0.012}$	0.306$_{\pm 0.015}$	+0.072	Vis
Paracentral lobule	**0.378**$_{\pm 0.021}$	0.324$_{\pm 0.011}$	+0.054	SomMot
Precentral gyrus medial	**0.369**$_{\pm 0.023}$	0.367$_{\pm 0.018}$	+0.002	SomMot
Superior occipital sulcus	**0.366**$_{\pm 0.013}$	0.296$_{\pm 0.014}$	+0.070	DorsAttn
Superior parietal lobule posterior	**0.364**$_{\pm 0.011}$	0.277$_{\pm 0.010}$	+0.087	DorsAttn
Parieto-occipital sulcus anterior	**0.362**$_{\pm 0.009}$	0.308$_{\pm 0.011}$	+0.054	DMN
Postcentral gyrus inferior	**0.361**$_{\pm 0.018}$	0.326$_{\pm 0.012}$	+0.035	SomMot
Occipital pole	**0.351**$_{\pm 0.015}$	0.280$_{\pm 0.010}$	+0.071	Vis
Precentral gyrus superior	**0.344**$_{\pm 0.018}$	0.293$_{\pm 0.018}$	+0.051	DorsAttn
Planum polare	0.343$_{\pm 0.021}$	**0.356**$_{\pm 0.013}$	−0.013	SalVAttn
Cingulate gyrus mid-anterior	0.326$_{\pm 0.021}$	**0.345**$_{\pm 0.017}$	−0.019	SalVAttn
Supramarginal gyrus	**0.323**$_{\pm 0.020}$	0.294$_{\pm 0.011}$	+0.029	SalVAttn
Middle frontal gyrus anterior	**0.321**$_{\pm 0.017}$	0.268$_{\pm 0.020}$	+0.053	SalVAttn
Superior temporal sulcus with angular gyrus	**0.306**$_{\pm 0.011}$	0.267$_{\pm 0.008}$	+0.039	DMN
Superior rostral gyrus	**0.300**$_{\pm 0.015}$	0.248$_{\pm 0.014}$	+0.052	DMN
Transverse sinus*	**0.300**$_{\pm 0.013}$	0.249$_{\pm 0.020}$	+0.051	Subcort
Parieto-occipital sulcus superior	**0.298**$_{\pm 0.011}$	0.228$_{\pm 0.016}$	+0.070	Cont
Inferior frontal sulcus	**0.298**$_{\pm 0.012}$	0.235$_{\pm 0.014}$	+0.063	Cont
Middle temporal gyrus	**0.294**$_{\pm 0.012}$	0.249$_{\pm 0.012}$	+0.045	DMN
Insula antero-superior	**0.292**$_{\pm 0.013}$	0.220$_{\pm 0.021}$	+0.072	SalVAttn
Inferior frontal gyrus	**0.291**$_{\pm 0.016}$	0.243$_{\pm 0.016}$	+0.048	Cont
Superior frontal sulcus	**0.287**$_{\pm 0.016}$	0.247$_{\pm 0.020}$	+0.040	DMN
Superior frontal gyrus medial	**0.286**$_{\pm 0.018}$	0.216$_{\pm 0.022}$	+0.070	SalVAttn
Dorsomedial prefrontal cortex antero-superior	**0.286**$_{\pm 0.013}$	0.235$_{\pm 0.021}$	+0.051	DMN
Cingulate cortex posterior	**0.284**$_{\pm 0.010}$	0.222$_{\pm 0.015}$	+0.062	DMN
Cingulate gyrus mid-posterior	**0.278**$_{\pm 0.010}$	0.230$_{\pm 0.015}$	+0.048	DMN
Middle frontal gyrus	**0.276**$_{\pm 0.011}$	0.218$_{\pm 0.014}$	+0.058	Cont
Intraparietal sulcus LH	**0.274**$_{\pm 0.012}$	0.220$_{\pm 0.015}$	+0.054	DorsAttn
Fusiform gyrus	0.274$_{\pm 0.015}$	**0.290**$_{\pm 0.014}$	−0.016	Cortical
Superior frontal gyrus	**0.265**$_{\pm 0.012}$	0.213$_{\pm 0.019}$	+0.052	DMN
Anterior Cingulate Cortex	**0.264**$_{\pm 0.012}$	0.244$_{\pm 0.017}$	+0.020	DMN
Angular gyrus inferior	**0.253**$_{\pm 0.008}$	0.221$_{\pm 0.014}$	+0.032	DMN
Intraparietal sulcus RH	**0.246**$_{\pm 0.012}$	0.189$_{\pm 0.016}$	+0.057	Cont
Cerebellum I-V	0.239$_{\pm 0.022}$	**0.297**$_{\pm 0.018}$	−0.058	Subcort
Intraparietal sulcus	**0.233**$_{\pm 0.012}$	0.185$_{\pm 0.015}$	+0.048	Cont
Thalamus	0.216$_{\pm 0.015}$	**0.317**$_{\pm 0.013}$	−0.101	Subcort
Superior fornix and isthmus*	0.207$_{\pm 0.025}$	**0.215**$_{\pm 0.019}$	−0.008	Subcort
Putamen	**0.205**$_{\pm 0.025}$	0.200$_{\pm 0.025}$	+0.005	Subcort
Angular gyrus superior	**0.195**$_{\pm 0.008}$	0.175$_{\pm 0.011}$	+0.020	DMN
Caudate	**0.186**$_{\pm 0.013}$	0.185$_{\pm 0.013}$	+0.001	Subcort
Cerebellum Crus II	0.143$_{\pm 0.012}$	**0.210**$_{\pm 0.018}$	−0.067	Subcort
ventricles*	0.026$_{\pm 0.015}$	**0.071**$_{\pm 0.011}$	−0.045	Subcort
All 64 ROIs	0.326	0.284	+0.042	
61 neural ROIs	0.333	0.289	+0.044	

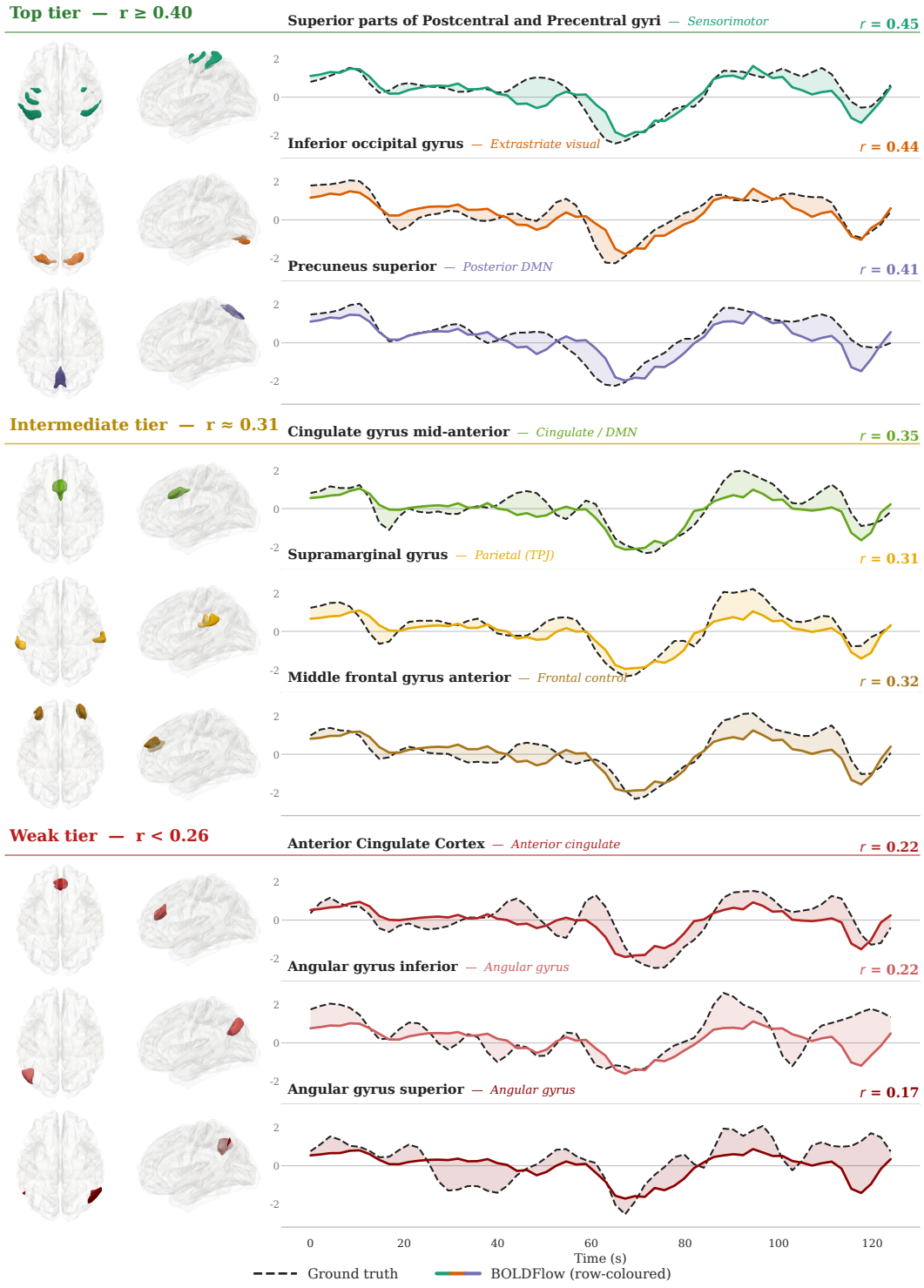

Figure 4: **Representative regional predictions on a held-out NeuroBOLT subject.** Solid lines show predictions and dashed lines measured BOLD; shading marks the difference between the curves. Nine DiFuMo components span higher and lower temporal correlations. The full trajectory-distribution audit is reported separately in Table 9.

G Covariance of Conditional Predictions

Proof of Proposition 1. For a BOLD vector $\mathbf{y}$ with finite second moment, EEG information $\mathcal{E}$, and a square-integrable predictor $f(\mathcal{E})$,

$$\mathbb{E}\|\mathbf{y} - f(\mathcal{E})\|^2 = \mathbb{E}\,\mathrm{tr}\,\mathrm{Cov}(\mathbf{y} \mid \mathcal{E}) + \mathbb{E}\|f(\mathcal{E}) - \mathbb{E}[\mathbf{y} \mid \mathcal{E}]\|^2.$$

The second term vanishes exactly at the conditional mean almost surely (Bishop, 2006). Applying the law of total covariance yields

$$\mathrm{Cov}(\mathbf{y}) = A + B, \qquad \mathrm{Cov}(f^\star(\mathcal{E})) = B, \qquad A \succeq 0.$$

For the same conditioning information $\mathcal{E}$, this conclusion holds componentwise: independently fitted per-region population MSE optima form the same vector of conditional means as the joint optimum. Joint fitting can change finite-sample estimation through shared inductive biases; both population optima retain covariance B.

With positive diagonal variances, correlation normalization gives $\mathcal{R}(B)$ and $\mathcal{R}(A+B)$. Provided their upper-triangle vectors have nonzero variance, their Pearson similarity is ρ_{CM}. By equality in Cauchy–Schwarz, $\rho_{\mathrm{CM}} = 1$ exactly when one centered vector is a positive multiple of the other, or equivalently when their uncentered vectors are related by a positive affine transformation. Consequently, $\rho_{\mathrm{CM}} < 1$ precisely when the two upper-triangle vectors are not related by a positive affine transformation. This condition concerns the pattern of normalized edge correlations, whereas $A \succeq 0$ describes the omitted covariance. □

Scope of the result. The proposition characterizes the exact population MSE optimum. Empirical regressors estimate that target with finite data and model capacity, while the large-ensemble readout estimates the fitted generative model's conditional mean. The experiments compare these fitted predictors directly.

The population in this decomposition must also be specified. Covariance pooled over participants includes between-participant effects that are absent from a within-scan covariance. For scan-level claims, the identities apply to the distribution relevant to that scan, or require assumptions connecting a pooled conditional model to it. Our empirical FC metric is computed within each scan.

Proof of Corollary 1. Let $\mathcal{E}$ collect the EEG windows contributing to an aggregated per-TR prediction. For a sampled output $\hat{\mathbf{y}}$, define

$$A_\Omega = \mathbb{E}_{\mathcal{E}}\,\mathrm{Cov}_\Omega(\hat{\mathbf{y}} \mid \mathcal{E}), \qquad B_\Omega = \mathrm{Cov}_{\mathcal{E}}(\mathbb{E}_\Omega[\hat{\mathbf{y}} \mid \mathcal{E}]).$$

The law of total covariance gives $\mathrm{Cov}(\hat{\mathbf{y}}) = A_\Omega + B_\Omega$. For M conditionally independent repetitions,

$$\mathrm{Cov}_\Omega\left(\frac{1}{M}\sum_{m=1}^{M}\hat{\mathbf{y}}^{(m)} \mid \mathcal{E}\right) = \frac{1}{M}\,\mathrm{Cov}_\Omega(\hat{\mathbf{y}} \mid \mathcal{E}),$$

while the conditional mean is unchanged. Applying total covariance again yields Eq. (6) exactly. For a stationary ergodic sequence with finite second moments, the ergodic theorem applied to first and second moments gives convergence of empirical covariance to the corresponding model marginal covariance. These convergence statements assume stationarity and ergodicity and allow temporal dependence. □

Overlap and conditional calibration. If K_t blocks contribute to a TR, its conditional covariance is

$$\mathrm{Cov}(\hat{\mathbf{y}}_t^{\mathrm{agg}} \mid \mathcal{E}) = \frac{1}{K_t^2}\sum_{k,\ell}\mathrm{Cov}(\hat{\mathbf{y}}_t^{(k)}, \hat{\mathbf{y}}_t^{(\ell)} \mid \mathcal{E}).$$

Independent anchor draws remove cross-anchor conditional terms, but the remaining variances depend on the block positions and EEG contexts. Averaging this conditional covariance over $\mathcal{E}$ gives A_Ω in Corollary 1. Matching the aggregated prediction's conditional first and second moments to those of $\mathbf{y}$ given the same $\mathcal{E}$ yields $(A_\Omega, B_\Omega) = (A, B)$.

Table 14: **Effect of trajectory averaging at fixed checkpoints.** BOLDFlow on NEUROBOLT, DiFuMo-64. Five-fold inter-subject CV with 3 seeds per fold (mean ± std across the 15 runs). At $M=1$, one sampled block per anchor is overlap-averaged to each TR. At $M>1$, the pointwise mean of M independently sampled trajectories is evaluated on all three metrics, including FC computed from that mean trajectory. EEG inputs are fixed throughout the sweep. NeuroBOLT joint-MSE regression provides the reference comparison. Best displayed values per metric are in bold, including ties.

Predictor	FC Corr. ↑	T. Corr. ↑	MSE ↓
Sampled-trajectory ($M=1$)	**0.584**$_{\pm 0.017}$	0.326$_{\pm 0.008}$	0.239$_{\pm 0.002}$
Ensemble-mean $M=5$	0.521$_{\pm 0.013}$	0.334$_{\pm 0.007}$	0.234$_{\pm 0.002}$
Ensemble-mean $M=10$	0.498$_{\pm 0.011}$	0.336$_{\pm 0.006}$	0.232$_{\pm 0.002}$
Ensemble-mean $M=25$	0.482$_{\pm 0.010}$	0.337$_{\pm 0.005}$	0.231$_{\pm 0.002}$
Ensemble-mean $M=50$	0.475$_{\pm 0.009}$	**0.338**$_{\pm 0.005}$	**0.230**$_{\pm 0.002}$
Ensemble-mean $M=200$	0.471$_{\pm 0.008}$	**0.338**$_{\pm 0.005}$	**0.230**$_{\pm 0.002}$
Reference: NeuroBOLT (joint MSE)	0.428$_{\pm 0.014}$	0.272$_{\pm 0.013}$	0.246$_{\pm 0.002}$

Prediction readout at fixed checkpoints. Table 14 holds model weights and EEG inputs fixed and averages sampled trajectories *before* computing FC, following the ensemble-mean estimator in Corollary 1.

From $M = 1$ to $M = 200$, T. Corr improves from 0.326 to 0.338 and MSE falls from 0.239 to 0.230, while FC Corr decreases from 0.584 to 0.471. Averaging attenuates model residual covariance by $1/M$, as characterized in Corollary 1. The observed trade-off favors sampled trajectories for connectivity and the ensemble mean for pointwise reconstruction. At $M = 200$, FC remains above joint-output NeuroBOLT's 0.428, so the trained model retains a connectivity advantage alongside the additional benefit of sampling.

Averaging FC estimates from separate trajectories is a different operation and does not algebraically remove the within-trajectory covariance in the same way. Likewise, pooling model first and second moments would define another estimator. Repeated draws can provide an uncertainty estimate through their spread while retaining the sampled FC readout. Appendices D and F provide the accompanying trajectory diagnostics and example reconstructions.

H EXTENDED RELATED WORK

Regional prediction and EEG representations. Regional EEG-to-BOLD models include interpretable approaches for deep-brain activity (Kovalev et al., 2022) and comparisons across classical and neural predictors (Donoso, 2025). EEG representation learning spans BENDR (Kostas et al., 2021), BIOT (Yang et al., 2023), LaBraM (Jiang et al., 2024), REVE (El Ouahidi et al., 2025), EEGPT (Wang et al., 2024), and CBraMod (Wang et al., 2025). These encoders provide learned EEG representations for downstream prediction.

Sources for conditional generation. Condition-specific anchors (Issachar et al., 2025) and Gaussian-process sources (Kollovieh et al., 2025) show that informative source distributions are part of a broader generative-modeling design space. Flow matching can transport fixed Gaussian sources as well (Lipman et al., 2023; Tong et al., 2024). BOLDFlow studies an EEG-conditioned Gaussian source, auxiliary moment-oriented supervision, and output readouts in a connectivity-recovery setting. Its ablations quantify the benefit of this configuration across the compared source and velocity designs.

Uncertainty interpretation. Predictive uncertainty is often discussed in terms of aleatoric and epistemic contributions (Kendall & Gal, 2017). Heteroscedastic training can couple mean and variance estimation (Seitzer et al., 2022; Stirn et al., 2023); β-NLL changes the weighting of those estimating equations. BOLDFlow's terminal ensemble spread is additionally affected by learned transport, finite paired data, and aggregation. We evaluate the resulting predictive spread through coverage, error ranking, and uncertainty-based edge selection.

I BROADER IMPACT AND LIMITATIONS

BOLDFlow addresses a methodological question: how well EEG can support recovery of a measured-fMRI connectome under paired-data supervision. Such prediction may eventually aid studies where repeated scanning is difficult, but synthetic BOLD remains an inference about another measurement modality. The present results establish neither replacement of fMRI nor clinical decision support.

Acquisition-domain scope. Both datasets contain simultaneous, in-scanner EEG from healthy adults. Outside-scanner recordings change device response, reference, montage, artifacts, and potentially behavioral state; coordinate encoding and normalization do not remove these shifts. Separate training on two datasets demonstrates replication, while acquisition-domain transfer requires prospective validation with a frozen protocol and matched-state recordings.

Conditional information and biological interpretation. EEG does not uniquely determine BOLD. Residual covariance may include unobserved neural activity, neurovascular variability, physiological fluctuations, motion, scanner effects, and preprocessing. Pairing and displacement controls show statistical input dependence but cannot isolate all of these sources. A fixed-residual-covariance comparator and nuisance-controlled replication would help determine what additional structure the learned conditional covariance captures.

Generated time series. Nonzero ensemble diversity, favorable FC, and recalibrated marginal intervals do not imply a calibrated BOLD process. The audit identifies compressed amplitudes, altered autocorrelation, and excess out-of-band power. Removing that excess leaves FC largely unchanged, but does not establish spectral, temporal, or physiological equivalence. The generated trajectories should be used within the evaluated connectivity-recovery scope.

Small cohorts and development selection. The 22- and 33-subject datasets remain small despite repeated scans and seeds. Subject-disjoint evaluation prevents direct subject overlap, but extensive development on fixed folds can still favor their validation structure. The fixed-configuration OpenNeuroSleep replication reduces dependence on one development setting; an independent, larger cohort would provide stronger confirmation. The seven-subject personalization experiment is exploratory, and its paired directions are not independent subjects.

Paired supervision and future validation. The decoder requires simultaneous paired EEG–fMRI during training; EEG-only pretraining does not remove this dependence. Pretraining a BOLD generative model on fMRI-only collections and using smaller paired sets for cross-modal alignment is a possible extension, not a demonstrated result. Task-state transfer, patient generalization, preservation of disease-related network effects, and outside-scanner use remain untested. Clinical claims would require dedicated cohorts and prospective validation of those endpoints.

Potential misuse. Predicted connectomes must remain distinguishable from measured or causally identified networks. Uncertainty-based selection offers a limited quality-control signal within the evaluated distribution; it does not protect against arbitrary distribution shifts.

Interpretation of network and uncertainty results. The covariance identities concern the population conditional mean and the effective aggregated model output; they do not bound arbitrary deterministic predictors or guarantee monotone FC similarity under averaging. The source ablations change learned scale, supervision, and coupling together. Native spread is input-conditional predictive variation rather than an identifiable pure aleatoric component; scalar recalibration evaluates marginal interval coverage, not the full conditional distribution. Fingerprinting between predicted scan halves can include session and model-fold signatures. Personalization exposes the adaptation-session connectome, so its improvement does not isolate the contribution of the test-session EEG. The 512-component atlas specifies output granularity, not 512 independently resolved neural sources; community recovery and edge selection were evaluated at the stated graph density and retention fraction.