SENSOR GEOMETRY AS A FLOW-MATCHING PRIOR FOR MULTI-CHANNEL BRAIN SIGNALS

Anonymous authors
Paper under double-blind review

ABSTRACT

Flow-matching models start from an isotropic Gaussian source, the standard choice when the correlation structure of the data is unknown in advance. For multi-channel brain recordings, however, part of this structure is known in advance. Electrodes sit at fixed positions on the head, and volume conduction through the skull and scalp makes nearby electrodes co-vary in a way that is shared across subjects. Existing EEG generative models nonetheless leave the network to learn this from scratch. We put this structure into the source instead. From the sensor coordinates alone, we build a k-nearest-neighbor graph and take a graph-Matérn function of its Laplacian as the source covariance, so the flow starts from spatially coherent patterns rather than channel-independent noise. The change adds no learned parameters, works with any coupling and any drift network, and uses the same three hyperparameters on every dataset. Across eight EEG datasets and four flow-matching methods, the graph-Matérn source lowers the spectral discrepancy between generated and real signals in the five clinical bands (PSD-KL) on most datasets. PSD-KL falls by 12% to 17% in geometric mean over datasets depending on the method and by up to 40% on PhysioNet-MI, the densest montage. We show that the improvement stems from the spatial eigenvectors of the local graph of sensor positions, since randomizing the eigenvectors while preserving the eigenvalue spectrum eliminates the gain. Furthermore, a prior fitted directly to the empirical data covariance performs worse than isotropic noise. The same construction applies unchanged to MEG, intracranial EEG with patient-specific grids, and a traffic-sensor network, lowering PSD-KL for every method on each[1].

1 INTRODUCTION

Flow-matching generative models (Lipman et al., 2023; Albergo et al., 2025; Liu et al., 2023; Tong et al., 2024a;b) start from an isotropic Gaussian, $x_0 \sim \mathcal{N}(0, \mathbf{I})$, and learn a drift that transports it to the data. For natural images this is the standard choice, since the correlation structure of the data is not known in advance. In multi-channel brain recordings, however, electrodes sit at fixed positions on the head, and volume conduction through skull and scalp makes nearby electrodes co-vary in a way that is shared across subjects (Nunez & Srinivasan, 2006). This structure is known before any data are seen, yet existing EEG generators, adversarial (Hartmann et al., 2018) and diffusion-based (Klein et al., 2024) alike, start from unstructured Gaussian noise and leave the sensor layout unused, so the network has to learn this structure from scratch.

We build this structure into the source instead, so that the flow starts from spatially coherent patterns and the drift does not have to learn the spatial correlations from scratch. Concretely, we connect each sensor to its k nearest neighbors, form the normalized Laplacian of that graph, and give each of its eigenvectors a variance set by a Matérn density of its eigenvalue (Borovitskiy et al., 2021) (see Figure 1). The resulting graph-Matérn prior puts most of its variance on spatial patterns that vary slowly across the scalp and little on patterns that flip sign between neighboring electrodes, so its samples are spatially coherent rather than channel-independent noise. This modification affects only the source distribution. It adds no learned parameters, works with any coupling and any drift network, and its three hyperparameters are fixed across all datasets.

[1] `https://anonymous.4open.science/r/graph-prior`

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.

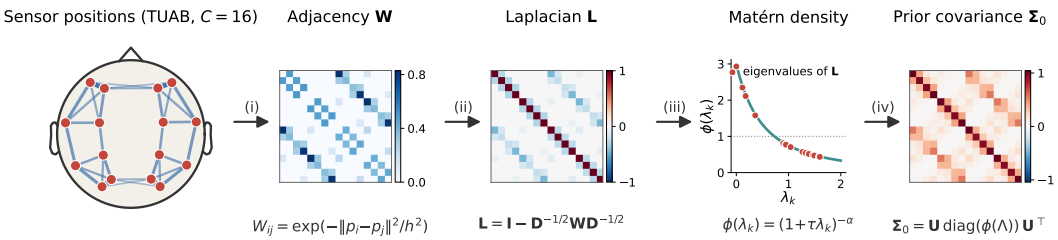

Figure 1: **Building the graph-Matérn source prior from sensor positions.** (i) Connect each sensor to its k nearest neighbors, with edge weights that decay with distance, to get the adjacency matrix $\mathbf{W}$. (ii) Form the normalized graph Laplacian $\mathbf{L}$, whose eigenvectors are spatial patterns over the sensors ordered from smooth to rough. (iii) Weight each pattern by the Matérn density $\phi(\lambda_m) = (1+\tau\lambda_m)^{-\alpha}$, which gives smooth patterns more variance and rough ones less. (iv) Assemble the resulting covariance $\boldsymbol{\Sigma}_0$, from which the flow's source samples x_0 are drawn in Eq. (2). Every step uses only the sensor coordinates and never the recorded signals.

We test the prior with four flow-matching methods, SF2M (Tong et al., 2024b), SI (Albergo et al., 2025), OT-CFM (Tong et al., 2024a), and RF (Liu et al., 2023), which cover the standard coupling choices, on eight EEG datasets, and measure how well generated signals reproduce the power spectrum of real signals in the five clinical frequency bands (PSD-KL; lower is better). We also apply it to MEG, to intracranial EEG with patient-specific electrode grids (van Blooijs et al., 2023), and to the PEMS-BAY traffic-sensor network (Li et al., 2018), to show that the prior is not specific to scalp EEG. Replacing the prior's eigenvectors, spectrum, or graph while keeping the other two fixed shows that the gain in PSD-KL comes from the eigenvectors of the local graph of sensor positions.

Our contributions are as follows.

- We propose a source prior for flow matching computed from sensor coordinates alone—a graph-Matérn covariance on the Laplacian of the graph of sensor positions—which replaces the standard isotropic source without adding learned parameters or modifying the drift network and coupling.
- We show that the prior lowers the spectral discrepancy (PSD-KL) for four flow-matching methods on most of the eight EEG datasets, by up to 40% on the densest montage, and that the same construction applies unchanged to MEG, intracranial EEG with patient-specific grids, and a traffic-sensor network.
- On TUAB, we show that the improvement stems from the spatial eigenvectors of the local position graph, and that a prior fitted directly to the empirical data covariance performs worse than isotropic noise.

2 Graph-Matérn Source Prior for Flow Matching

2.1 Preliminaries

Let p_0 and p_1 be the prior and data distributions on $\mathbb{R}^{C\times T}$ (channels $\times$ time), and $\pi \in \Pi(p_0, p_1)$ be a coupling. Flow matching (Lipman et al., 2023) trains a neural vector field $v_\theta(x, t)$ to match the conditional velocity $u_t(x_t \mid x_0, x_1) = \partial_t I(t, x_0, x_1)$ along an interpolation path parameterized by $t \sim \mathcal{U}(0, 1)$; for the straight line $I(t, x_0, x_1) = (1-t)x_0 + tx_1$ used in OT-CFM (Tong et al., 2024a) this reduces to $u_t = x_1 - x_0$, giving the squared-error loss of Eq. (1).

$$\mathcal{L}_{\mathrm{FM}} = \mathbb{E}_{t\sim\mathcal{U}(0,1),\,(x_0,x_1)\sim\pi,\,x_t=I(t,x_0,x_1)} \left\| v_\theta(x_t, t) - u_t \right\|_2^2. \tag{1}$$

Stochastic interpolants (Albergo et al., 2025) add a latent Gaussian noise term, yielding $X_t = (1-t)x_0 + tx_1 + \gamma(t)z$ with $z \sim \mathcal{N}(0, \mathbf{I})$ and $\gamma(0) = \gamma(1) = 0$ so that $X_0 = x_0$, $X_1 = x_1$. Differentiating in t gives $u_t = (x_1 - x_0) + \gamma'(t)z$. The marginal drift $b(x, t) = \mathbb{E}[u_t \mid X_t = x]$ is the probability-flow ODE drift for the interpolant law (Albergo et al., 2025). Finally, for a graph of sensor positions $G = (\mathcal{V}, \mathcal{E})$ with $|\mathcal{V}| = C$, adjacency matrix $\mathbf{W}$, and degree matrix $\mathbf{D}$, we denote the symmetric normalized Laplacian by $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$, with eigenvalues $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_C)$ where $0 \le \lambda_m \le 2$.

Table 1: Channel-covariance rank deficiency per dataset with C channels. Top-k columns give the fraction of the total variance $\operatorname{tr}\boldsymbol{\Sigma}_1$ that lies along the top-k eigenvectors of the channel covariance $\boldsymbol{\Sigma}_1$; $d_{\text{eff}} = (\operatorname{tr}\boldsymbol{\Sigma}_1)^2/\operatorname{tr}(\boldsymbol{\Sigma}_1^2)$ is the participation-ratio effective dimension. The covariance is estimated over all time samples of 512 windows of the test split, centered per channel.

dataset	# channels (C)	top-1	top-3	$d_{\text{eff}}(\boldsymbol{\Sigma}_1)$
TUAB (Obeid & Picone, 2016)	16	0.387	0.647	4.95
TUEV (Obeid & Picone, 2016)	16	0.330	0.605	5.87
Mumtaz-MDD (Mumtaz, 2016)	19	0.997	0.999	1.01
BCI-IV 2a (Brunner et al., 2008)	22	0.494	0.855	3.02
FACED (Chen et al., 2023)	32	0.239	0.607	7.26
SHU (Ma et al., 2022)	32	0.576	0.816	2.75
SEED-V (Liu et al., 2022)	62	0.549	0.796	2.98
PhysioNet-MI (Schalk et al., 2004)	64	0.635	0.775	2.40

2.2 Definition

Motivation. The channel covariance of a brain recording is shaped in part by the sensor layout. For example, in scalp EEG, volume conduction and shared cortical generators couple nearby electrodes (Nunez & Srinivasan, 2006), and local cortical structure does the same for intracranial recordings. Empirically, this makes the covariance matrix effectively low-rank. In the eight EEG datasets we evaluate, the top-3 eigenvectors explain between 61% and 100% of the total channel variance, and the participation-ratio effective dimension is between 1.0 and 7.3 for 16 to 64 channels (see Table 1). An isotropic source ignores this structure, requiring the drift network to transport white noise to an effectively low-rank target. As shown in Section 4, using the empirical data covariance directly degrades performance below the isotropic baseline, whereas aligning with the eigenvectors of the position graph drives the performance improvement.

Construction. Given 3-D sensor coordinates $\{p_c\}_{c=1}^{C}$, we build an undirected k-nearest-neighbor graph without self-loops (W_{ii}=0), connecting sensors whenever either is among the other's k nearest neighbors. Edge weights are set to $W_{ij} = \exp(-\|p_i - p_j\|^2/h^2)$, with bandwidth h^2 defined as the mean squared k-NN distance across all sensors. We then compute the normalized Laplacian $\mathbf{L} = \mathbf{I} - \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top$ with eigenvalues $0 = \lambda_1 \le \lambda_2 \le \cdots \le \lambda_C \le 2$ (see Figure 1). We place an independent Gaussian on each eigenmode with variance $\phi(\lambda_m) = (1 + \tau\lambda_m)^{-\alpha}$ and reassemble in sensor coordinates,

$$\operatorname{vec}(x_0) \;=\; (\mathbf{I}_T \otimes \mathbf{U})\operatorname{vec}(\tilde{x}_0), \qquad \tilde{x}_0 \sim \mathcal{N}\big(0,\, \mathbf{I}_T \otimes \operatorname{diag}(\phi(\boldsymbol{\Lambda}))\big), \tag{2}$$

where $x_0, \tilde{x}_0 \in \mathbb{R}^{C\times T}$ and $\otimes$ denotes the Kronecker product. Equivalently, $x_0 = \mathbf{U}\,\tilde{x}_0$ left-multiplies each time-slice by $\mathbf{U}$. The $\mathbf{I}_T$ in the temporal slot is a deliberate simplification, since we encode no temporal structure in the prior and leave any temporal correlations for the drift network to learn from data. The scalar $\phi(\lambda)$ is the graph-Matérn spectral density of Borovitskiy et al. (2021), a discrete-graph analog of the Matérn Gaussian process density on Riemannian manifolds (Borovitskiy et al., 2020), where τ sets the spatial correlation length and α the spectral decay rate. We normalize ϕ so that $\frac{1}{C}\sum_m \phi(\lambda_m) = 1$, placing the prior on the same total-variance scale as an isotropic source; the prior *redistributes* variance across spatial modes. The prior has three hyperparameters, k, τ, and α, whose experimental settings are described in Section 3.1.

Properties. The prior aligns the source variance with the eigenvectors of the Laplacian of the position graph and concentrates it on the low-eigenvalue modes, lowering the prior's effective dimension $d_{\text{eff}}(\boldsymbol{\Sigma}_0) := (\operatorname{tr}\boldsymbol{\Sigma}_0)^2/\operatorname{tr}(\boldsymbol{\Sigma}_0^2)$ from 16 to 10.1 on TUAB[2]. When evaluated using the Tikhonov effective dimension $\operatorname{tr}(\boldsymbol{\Sigma}_0(\boldsymbol{\Sigma}_0 + \lambda\mathbf{I})^{-1})$, which discounts small eigenvalues, the reduction is at most 10% (see Appendix C); as shown in Section 4, the spatial eigenvectors, rather than variance concentration, drive the empirical improvement.

[2]This is the *prior*'s effective dimension, distinct from the data-covariance $d_{\text{eff}}(\boldsymbol{\Sigma}_1)$ in Table 1.

Algorithm 1: Training with the graph-Matérn source prior

Input: dataset $\mathcal{D} = \{x_1^{(i)}\}$, sensor coordinates $\{p_c\}_{c=1}^{C}$, prior hyperparameters (τ, α, k), noise scale σ with $\gamma(t) = \sigma\sqrt{t(1-t)}$, ϵ=0.001, coupling π, drift network v_θ, learning rate η, number of steps N.
Output: Learned drift $v_\theta^\star$.
Build k-NN adjacency $\mathbf{W}$ with Gaussian weights $W_{ij} = \exp(-\|p_i - p_j\|^2/h^2)$, h^2 the mean squared k-NN distance, symmetrize, compute normalized Laplacian $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$; precompute $\phi(\lambda_m) = (1+\tau\lambda_m)^{-\alpha}$ normalized so $\frac{1}{C}\sum_m \phi(\lambda_m) = 1$;
for *step* $= 1, \ldots, N$ **do**
  Sample data batch $\{x_1^{(i)}\} \sim \mathcal{D}$;
  Sample prior batch $x_0 = \mathbf{U}\,\mathrm{diag}(\sqrt{\phi(\mathbf{\Lambda})})\, z_0$, $z_0 \sim \mathcal{N}(0, \mathbf{I}_{C\times T})$; `// only change`
  Couple $(x_0, x_1) \sim \pi$;
  Sample $t \sim \mathcal{U}(\epsilon, 1-\epsilon)$, $z \sim \mathcal{N}(0, \mathbf{I})$; form $X_t = (1-t)x_0 + tx_1 + \gamma(t)z$, $u_t = (x_1 - x_0) + \gamma'(t)z$;
  Update $\theta \leftarrow \theta - \eta\,\nabla_\theta \|v_\theta(X_t, t) - u_t\|_2^2$;
end

2.3 Composition with Flow-Matching Couplings

The graph-Matérn prior is independent of the coupling $\pi(x_0, x_1)$ and of the drift architecture. The four methods we evaluate are listed in Section 3. The training objective is the loss of Eq. (1) with the stochastic-interpolant target $u_t = (x_1 - x_0) + \gamma'(t)z$ of Section 2.1, with $\gamma(t) = \sigma\sqrt{t(1-t)}$, the vector field evaluated at the stochastic state X_t, and the expectation taken over z as well. Relative to an isotropic baseline, the only change is the source distribution p_0 (the marked line in Algorithm 1). At inference, X_0 is drawn from the prior and the learned drift is integrated with 50 Euler steps of the probability-flow ODE across all methods. Consequently, the stochastic formulations (SI (Albergo et al., 2025)[3] and SF2M (Tong et al., 2024b), with σ>0) are trained with interpolant noise and sampled deterministically without score networks (see Appendix A), sharing the ODE integration scheme with the deterministic methods (OT-CFM (Pooladian et al., 2023; Tong et al., 2024a), RF (Liu et al., 2023)).

3 Experiments

3.1 Experimental Details

Methods. We test four flow-matching methods, each with an isotropic source ($\mathcal{N}(0, \mathbf{I})$) and with the graph-Matérn source of Section 2.2, namely SF2M (Tong et al., 2024b) (Hungarian coupling, σ>0)[4], SI (Albergo et al., 2025) (independent coupling, σ>0), OT-CFM (Pooladian et al., 2023; Tong et al., 2024a) (Hungarian coupling, σ=0), and RF (Liu et al., 2023) (independent coupling, σ=0, followed by one reflow). A method with the graph prior carries the prefix *GP-* (e.g., GP-SI is SI with the graph-Matérn source). All methods share a 1D U-Net drift (Ho et al., 2020; Ronneberger et al., 2015) that convolves along time and treats the C sensors as input channels, so the network has no spatial inductive bias and the sensor layout enters only through the source; architecture, sweep configuration, and sampler details are in Appendix A.

Hyperparameters. We fix the prior hyperparameters to k=4, τ=1, and α=2 across all datasets (sensitivity analysis in Appendix D). For methods with a stochastic bridge (σ>0), we use noise scale σ=0.02, adjusted to σ=0.008 on TUEV to account for its 2.5× smaller signal amplitude, which maintains $\sigma/\sigma_{\text{real}} \approx 0.08$ relative to the per-channel training standard deviation σ_{real} on every dataset except FACED, where it is 0.19 (see Appendix E).

[3]We abbreviate the σ>0 *stochastic interpolant* of Albergo et al. (2025) under independent coupling as SI throughout the paper.

[4]We use the exact minibatch-OT (Hungarian) coupling on each batch of 64 (SF2M-Exact in Tong et al. (2024b)), so that SF2M and OT-CFM share the coupling and differ only in σ.

Table 2: PSD-KL on eight EEG datasets (test split, mean±std over five training seeds; lower is better) for the four flow-matching methods with an isotropic source and with the graph-Matérn source (GP-). Bold denotes the better source within each method. Because PSD-KL is not comparable across datasets, the last row gives the geometric mean over datasets of the ratio GP/isotropic (below one favors the graph source).

	σ>0 Hungarian		σ>0 independent		σ=0 Hungarian		σ=0 indep+reflow	
dataset (ch)	SF2M	GP-SF2M	SI	GP-SI	OT-CFM	GP-OT-CFM	RF	GP-RF
TUAB (16)	1.68±0.23	**1.04±0.28**	1.43±0.37	**1.16±0.20**	1.67±0.33	**1.40±0.38**	**2.27±1.70**	2.53±1.13
TUEV (16)	**2.03±0.33**	3.85±2.47	2.55±0.78	**2.15±1.10**	5.30±2.30	**3.80±0.59**	3.19±1.2	**3.09±0.29**
Mumtaz (19)	4.42±3.33	**3.89±1.94**	3.91±2.42	3.91±1.11	**4.97±2.06**	5.51±1.66	6.06±3.0	**3.79±1.3**
BCI-IV 2a (22)	3.78±1.00	**3.56±1.47**	**3.30±0.73**	3.50±0.59	**3.06±0.73**	3.61±0.19	4.27±0.71	**4.24±1.3**
FACED (32)	32.9±4.0	**31.8±7.8**	33.3±3.5	**29.0±4.6**	33.6±7.2	**31.5±12.0**	32.1±3.9	**26.4±5.0**
SHU (32)	113±3.9	**81.4±8.0**	114±1.7	**82.8±4.6**	111±5.2	**85.4±0.7**	116±2.3	**82.9±10.3**
SEED-V (62)	32.8±0.7	**28.1±0.2**	32.9±0.7	**28.0±0.1**	32.9±0.7	**28.1±0.3**	33.3±0.4	**28.3±0.2**
PhysioNet-MI (64)	34.8±0.9	**20.7±3.7**	33.4±0.9	**19.9±2.4**	33.0±1.4	**19.9±3.9**	33.3±1.8	**22.4±1.5**
geometric mean of GP/iso	0.88		0.83		0.86		0.83	

Metrics. Our primary metric, PSD-KL, fits a Gaussian to the log band power across windows for each channel and clinical band ($\delta, \theta, \alpha, \beta, \gamma$) and reports the symmetric KL divergence between the real and generated fits, averaged over channels and bands (lower is better; see Appendix A). Because the divergence scales with each dataset's empirical variance of log band power and diverges under variance mismatch, absolute values are not comparable across datasets; comparisons are made within each dataset. Band power is the standard summary of an EEG recording (Nunez & Srinivasan, 2006), whereas general-purpose scores like the Fréchet Inception Distance (Heusel et al., 2017) can disagree with spectral fidelity on EEG, where the model with the most realistic spatial and spectral properties was assigned the worst FID (Hartmann et al., 2018). To complement this per-channel metric, we quantify cross-channel coupling using the weighted phase-lag index (wPLI) (Vinck et al., 2011), evaluated as the Pearson correlation between real and generated wPLI matrices (higher is better).

Data and statistics. We evaluate on eight EEG datasets (TUAB and TUEV (Obeid & Picone, 2016), Mumtaz-MDD (Mumtaz, 2016), BCI-IV 2a (Brunner et al., 2008), FACED (Chen et al., 2023), SHU (Ma et al., 2022), SEED-V (Liu et al., 2022), and PhysioNet-MI (Schalk et al., 2004)) spanning 16 to 64 channels across resting-state, event-related, motor-imagery, and emotion paradigms. Preprocessing and data splits follow Wang et al. (2025). All metrics are evaluated on the official test split, which holds out subjects on six of the eight datasets (SEED-V partitions trials within session, while SHU shares sessions across splits; see Appendix A). All experiments use the same five training seeds, shared across both sources; reported values indicate the mean ± standard deviation over seeds. The MEG, intracranial EEG, and traffic experiments of Section 3.3 use the same methods, sources, seeds, and metrics; their graphs, training lengths, and splits are in Appendix G.

3.2 The Graph Prior Improves the Power Spectrum on EEG

Power spectrum. As shown in Table 2, the graph prior lowers PSD-KL across all four flow-matching formulations, achieving geometric mean ratios of 0.83–0.88 relative to the isotropic baseline. For every method, it lowers PSD-KL on most datasets, and the exceptions are discussed at the end of this section. The largest reductions shared by all four methods occur on PhysioNet-MI (64 channels), where PSD-KL drops from 33–35 to 20–22, and on SHU (32 channels), where it drops from 111–116 to 81–85. A one-sided Wilcoxon signed-rank test on $\log(\text{iso}/\text{graph})$ across the eight datasets yields uncorrected p-values of 0.016 for SI, 0.020 for RF, 0.055 for OT-CFM, and 0.098 for SF2M (adjusted to 0.063 for SI and RF after Holm correction). Figure 2 illustrates these qualitative differences on TUAB. GP-SI samples retain the slow oscillations and inter-channel coherence of real EEG, whereas SI samples resemble channel-independent high-frequency noise.

Cross-channel coupling. The graph prior raises the wPLI correlation on average for all four methods by 0.01–0.03 as shown in Table 3, with consistent gains across all four methods on the densest montages, SEED-V and PhysioNet-MI (+0.05–0.12). On the remaining datasets, differences generally remain within standard deviations, except on TUAB, where GP-SI improves and GP-OT-CFM decreases by 0.06, and on FACED, where GP-RF decreases by 0.05. Across the eight datasets,

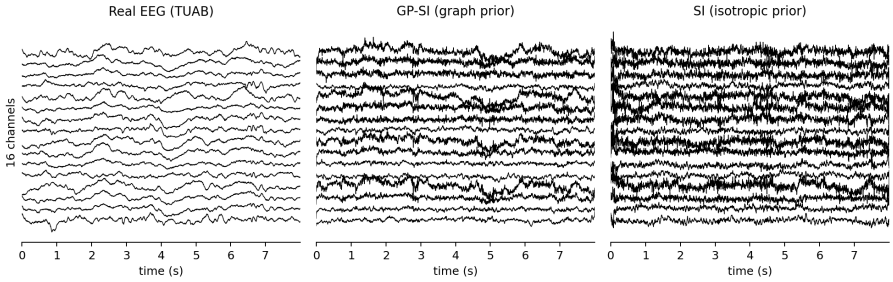

Figure 2: **Qualitative effect of the graph prior (GP) on TUAB (Obeid & Picone, 2016).** 16-channel waveforms from one real test window and one sample each from GP-SI (SI trained with the graph prior) and SI (Albergo et al., 2025). Real EEG shows slow oscillations that are coherent across channels; GP-SI samples retain both properties, whereas SI samples resemble channel-independent high-frequency noise.

Table 3: wPLI correlation between real and generated data on the same eight datasets (band-averaged correlation, mean±std over five training seeds; higher is better) for the four methods with an isotropic and with the graph-Matérn source. Bold denotes the better source within each method.

	$\sigma>0$ Hungarian		$\sigma>0$ independent		$\sigma=0$ Hungarian		$\sigma=0$ indep+reflow	
dataset (ch)	SF2M	GP-SF2M	SI	GP-SI	OT-CFM	GP-OT-CFM	RF	GP-RF
TUAB (16)	$\mathbf{0.11\pm0.10}$	0.06 ± 0.12	0.02 ± 0.05	$\mathbf{0.08\pm0.05}$	$\mathbf{0.13\pm0.05}$	0.07 ± 0.05	0.03 ± 0.04	$\mathbf{0.05\pm0.05}$
TUEV (16)	-0.00 ± 0.03	$\mathbf{0.01\pm0.04}$	$\mathbf{0.03\pm0.02}$	0.00 ± 0.05	-0.03 ± 0.04	$\mathbf{-0.01\pm0.07}$	0.01 ± 0.02	0.01 ± 0.05
Mumtaz (19)	0.31 ± 0.16	$\mathbf{0.41\pm0.06}$	$\mathbf{0.40\pm0.03}$	0.39 ± 0.02	0.39 ± 0.06	$\mathbf{0.40\pm0.03}$	0.33 ± 0.13	$\mathbf{0.38\pm0.07}$
BCI-IV 2a (22)	0.19 ± 0.05	$\mathbf{0.22\pm0.07}$	0.19 ± 0.08	$\mathbf{0.22\pm0.06}$	0.17 ± 0.08	$\mathbf{0.23\pm0.06}$	$\mathbf{0.17\pm0.04}$	0.13 ± 0.03
FACED (32)	0.04 ± 0.05	$\mathbf{0.06\pm0.07}$	0.08 ± 0.03	$\mathbf{0.09\pm0.06}$	0.06 ± 0.03	$\mathbf{0.10\pm0.05}$	$\mathbf{0.12\pm0.02}$	0.07 ± 0.02
SHU (32)	$\mathbf{0.03\pm0.04}$	0.01 ± 0.04	$\mathbf{0.02\pm0.04}$	0.01 ± 0.02	$\mathbf{0.03\pm0.05}$	0.01 ± 0.02	0.01 ± 0.01	$\mathbf{0.02\pm0.03}$
SEED-V (62)	-0.01 ± 0.01	$\mathbf{0.06\pm0.02}$	-0.01 ± 0.02	$\mathbf{0.06\pm0.01}$	-0.01 ± 0.02	$\mathbf{0.06\pm0.03}$	-0.02 ± 0.02	$\mathbf{0.03\pm0.04}$
PhysioNet-MI (64)	0.33 ± 0.05	$\mathbf{0.40\pm0.04}$	0.32 ± 0.05	$\mathbf{0.44\pm0.02}$	0.30 ± 0.04	$\mathbf{0.42\pm0.06}$	0.29 ± 0.03	$\mathbf{0.38\pm0.06}$
mean across datasets	0.12	**0.15**	0.13	**0.16**	0.13	**0.16**	0.12	**0.13**

these differences do not reach statistical significance after Holm correction. Because the graph prior directly shapes zero-lag spatial covariance, whereas wPLI explicitly discards zero-lag interactions to eliminate volume conduction, the small average effect on this metric (0.01 to 0.03) is expected.

Datasets without a gain. Each case in Table 2 where the graph prior does not improve on isotropic noise has a likely explanation. On BCI-IV 2a, the 22 electrodes cover only the centro-parietal region, so the graph has little spatial structure to encode, whereas a similar motor-imagery task recorded with 64 electrodes over the whole scalp (PhysioNet-MI) gives our largest overall gain. On FACED, the prior lowers PSD-KL for all four methods at the default noise scale, and only at a five times larger one (σ=0.10) is it worse in β and γ while matching or improving the lower bands. However, because the prior lowers the γ-band error on SHU from 340 to 261 for SI, this behavior likely reflects dataset-specific high-frequency noise rather than an inherent limit of spatial priors. On TUEV, only SF2M is worse with the prior, and its increase is spread over all five frequency bands. On Mumtaz, SI is unchanged and OT-CFM is 11% worse, and on TUAB, RF is 11% worse, with these differences remaining within seed standard deviations. Appendix H gives the per-band results behind the FACED, SHU, and TUEV statements.

3.3 Generalization to MEG, Intracranial EEG, and Traffic Sensors

To test whether the graph prior applies beyond multi-channel scalp EEG, we evaluate the identical construction on other spatial sensor systems using coordinates alone (see Appendix G). As shown in Table 4, the graph prior lowers PSD-KL across all three domains, whole-head MEG (Wakeman–Henson (Wakeman & Henson, 2015); 16 subjects, 102 magnetometers on a shared helmet), intracranial EEG (CCEP ECoG (van Blooijs et al., 2023); five patients with individual 72–112-channel ECoG grids), and the PEMS-BAY traffic network (Li et al., 2018) (325 freeway loop sensors, graph from GPS coordinates).

The relative reduction is largest on iEEG (34–40%), where each patient has a distinct cortical grid, intermediate on the traffic network (11%), and smallest on MEG (6%), where all subjects share a

Table 4: PSD-KL on MEG, intracranial EEG, and a traffic-sensor network (lower is better; mean over five seeds and, for MEG and iEEG, over subjects or patients) for the four methods with an isotropic and with the graph-Matérn source. Bold denotes the better source within each method. The graph prior lowers PSD-KL across all four methods and all three modalities (one-sided sign test, $p < 0.0001$ over MEG subjects and $p = 0.031$ over iEEG patients and over traffic seeds). Absolute values are not comparable across rows; per-patient iEEG values are in Table 13.

	σ>0 Hungarian		σ>0 independent		σ=0 Hungarian		σ=0 indep+reflow	
modality	SF2M	GP-SF2M	SI	GP-SI	OT-CFM	GP-OT-CFM	RF	GP-RF
MEG, Wakeman–Henson (16 subjects)	2890	**2715**	2891	**2715**	2893	**2712**	2902	**2726**
iEEG, CCEP ECoG (5 patients)	31.3	**19.2**	31.2	**18.7**	30.5	**18.9**	37.6	**24.9**
PEMS-BAY traffic (325 sensors)	62.1	**55.6**	62.1	**55.5**	62.1	**55.3**	63.9	**56.8**

Table 5: Downstream augmentation on Mumtaz-MDD with an EEGNet-8,2 classifier (subject-disjoint split with 31 training, 6 validation, and 8 test subjects; every column is the mean±std over five seeds). Rows add one generated window per real training window from the isotropic or the graph-prior source, each generated window first matched in amplitude to the real training windows (see Appendix A). Bold denotes the better source.

method	balanced accuracy	recall, minority class	recall, majority class
real only	0.608 ± 0.146	0.284 ± 0.386	0.931 ± 0.096
class-weighted real only	0.545 ± 0.097	0.091 ± 0.194	0.999 ± 0.002
+ iso-source	0.819 ± 0.037	0.786 ± 0.009	0.852 ± 0.075
+ graph-prior	$\mathbf{0.833 \pm 0.054}$	$\mathbf{0.794 \pm 0.006}$	$\mathbf{0.872 \pm 0.110}$

fixed sensor geometry. Unlike neural recordings, the freeway network involves no biological tissue or volume conduction. The improvement on this domain indicates that the prior's benefit stems from the spatial geometry encoded in the graph eigenvectors (analyzed in Section 4), rather than domain-specific electrophysiological properties.

3.4 Data Augmentation for EEG Classification

To evaluate whether generative improvements translate to downstream utility, we benchmark data augmentation on Mumtaz-MDD, a binary depression-detection task with a subject-disjoint split (31 training, 6 validation, and 8 test subjects, 841 test windows). We train class-conditional generators on the training split only and train an EEGNet-8,2 classifier (Lawhern et al., 2018) on real windows with and without synthetic data added. Before augmenting the training set, generated windows are matched in root-mean-square amplitude to the real training data (see Appendix A). As shown in Table 5, synthetic windows increase balanced accuracy from 60.8% to 83.3% with the graph prior and to 81.9% with the isotropic baseline. This gain is driven primarily by the minority class, where recall rises from 0.284 to 0.794. The improvement is not achievable by loss reweighting alone, as class-weighted training on real data reaches only 54.5% balanced accuracy.

4 Which Part of the Graph Prior Carries the Gain?

We next analyze which properties of the graph prior account for its empirical gains. The graph-Matérn source differs from isotropic noise in three structural aspects. Its variance is concentrated on fewer modes (effective dimension 10.1 vs. 16.0 on TUAB), its spectrum decays smoothly along the Laplacian eigenvalues, and its eigenvectors are defined by the graph of sensor positions. Any of these factors could explain the observed improvement. We therefore train SI on TUAB, the simplest of the four methods, with priors that replace the eigenvectors, the spectrum, or the graph while keeping the rest fixed, and compare them in Table 6 against the isotropic source (PSD-KL 1.85 ± 0.18) and the graph prior (1.08 ± 0.41).

The gain requires the physical sensor eigenbasis. To test whether the improvement stems from sensor geometry or merely from low effective dimension, we keep the Matérn spectrum (d_{eff}=10.1) but substitute the eigenvectors (see Appendix F.1). Replacing them with permuted sensor positions (3.70 ± 2.79 over 25 permutations), a uniformly random orthonormal basis (4.42 ± 2.97), or empirical covariance eigenvectors (2.76 ± 2.93) is worse than the isotropic baseline (1.85 ± 0.18) and raises

Table 6: Ablating the graph-Matérn prior on TUAB with SI. PSD-KL on the validation split at the configuration of TUAB's row in Table 2, mean±std over five training seeds, over 25 random permutations for the shuffled-position row, and over ten seeds for the correlation-graph row; the reference rows therefore differ from the test-split values of that row. The second column names the part of the construction each row replaces, d_{eff} is the prior's effective dimension, and bold denotes the best prior (details in Appendix F).

prior	changes	d_{eff}	PSD-KL
isotropic	—	16.0	1.85 ± 0.18
graph-Matérn (graph of sensor positions)	—	10.1	$\mathbf{1.08 \pm 0.41}$
shuffled sensor positions	eigenvectors	10.1	3.70 ± 2.79
uniformly random eigenvectors	eigenvectors	10.1	4.42 ± 2.97
empirical eigenvectors, graph spectrum	eigenvectors	10.1	2.76 ± 2.93
empirical covariance	eigenvectors and spectrum	5.2	4.61 ± 2.40
Ledoit–Wolf-shrunk empirical covariance	eigenvectors and spectrum	6.6	3.56 ± 2.43
heat-kernel spectrum	spectrum	12.0	1.38 ± 0.86
hard low-pass spectrum	spectrum	8.0	71.4 ± 12.5
fully connected Gaussian graph	graph	11.2	2.53 ± 1.81
correlation graph (k-NN on $\lvert\text{corr}\rvert$)	graph	11.0	1.39 ± 1.70

the standard deviation over seeds about sevenfold relative to the graph prior (0.41). This failure illustrates that an isotropic prior is at least spatially neutral; an eigenbasis other than the graph's, by contrast, imposes correlations that the model must undo during integration. Spectral concentration is therefore beneficial only when guided by the physical sensor layout.

Concentrating variance in the wrong directions is worse than no structure. Using the empirical channel covariance itself as the prior combines eigenvectors estimated from the data with a more concentrated spectrum (d_{eff}=5.2). At 4.61 ± 2.40, this prior underperforms all priors that replace only the eigenvectors, demonstrating that concentrating variance is unhelpful unless aligned with the eigenvectors of the position graph. Shrinking the empirical covariance toward a multiple of the identity with the Ledoit–Wolf estimator (Ledoit & Wolf, 2004), which keeps its eigenvectors and only flattens its spectrum, is also worse than the isotropic source (3.56 ± 2.43 against 1.85 ± 0.18; see Appendix F.2). This failure is not caused by near-zero eigenvalues in the empirical covariance, since the prior that keeps empirical eigenvectors but uses the graph's Matérn spectrum, which has no small eigenvalues, also underperforms the isotropic source (2.76 ± 2.93). Nor can the failure be explained by Gaussian source–target covariance mismatch, as on a synthetic Gaussian target matching the EEG channel covariance, all six tested Gaussian sources reach a similar PSD-KL (0.0020–0.0032; see Appendix F.2). The gap on real EEG therefore comes from its non-Gaussianity and its temporal correlations, which this Gaussian target lacks.

The spectrum only needs to be smooth and to cover every mode. Replacing the Matérn density with a heat kernel $\exp(-\tau\lambda)$ on the same eigenvectors keeps most of the gain (1.38 ± 0.86; see Appendix F.3). A hard low-pass that zeroes the upper half of the modes fails (71.4 ± 12.5), since the flow is a smooth invertible map and cannot spread a source with no variance on half of the modes over a target that has variance on all of them; the drift has to create all of the missing variance.

The graph must be sparse and local. Connecting all sensor pairs with a fully connected Gaussian graph rather than restricting edges to k nearest neighbors yields 2.53 ± 1.81, performing worse than isotropic noise (1.85 ± 0.18). Sensor geometry alone is therefore insufficient; the inductive bias requires the eigenvectors of a sparse local graph. A k-NN graph built from channel correlations instead of positions yields 1.39 ± 1.70 across ten seeds (median 0.90); while it is competitive on most seeds, its standard deviation over seeds is four times that of the graph prior, likely because data-driven correlations inherit recording artifacts absent in physical coordinates (see Appendix F.4). In summary, these experiments indicate that the gain is carried by the eigenvectors of a sparse local graph over the sensors. The spectrum needs only to be smooth and full-rank, and sensor coordinates provide such a graph without access to the data. Two baselines that instead whiten the data with the empirical channel covariance (ZCA whitening) or put the graph into the drift network (a graph-convolution mixer) perform within the standard deviation of the isotropic baseline at this configuration (see Appendix F.5).

5 Related Work

Structured sources and geometric priors in generative models. Non-isotropic noise schedules in diffusion alter the covariance profile along the sampling trajectory while retaining an isotropic base distribution (Dockhorn et al., 2022; Hoogeboom & Salimans, 2023; Shaul et al., 2023). Similarly, geometric generative models build physical constraints into the drift or score network (De Bortoli et al., 2022; Dutordoir et al., 2023; Hoogeboom et al., 2022). Few methods change the base distribution itself. PriorGrad (Lee et al., 2022) adapts diagonal covariance matrices to conditioning inputs in speech synthesis, non-isotropic (Voleti et al., 2022) and function-space diffusions (Kerrigan et al., 2023; Pidstrigach et al., 2024) define Gaussian-process priors over continuous domains, and graph-aware diffusion (Rozada et al., 2026) obtains a graph-Laplacian covariance from heat diffusion in the forward process. Within flow matching, TSFlow (Kollovieh et al., 2025) structures the source along the temporal axis with a Gaussian process for time-series forecasting. In contrast, our approach leaves the flow-matching objective, couplings, and drift parameterization unchanged and uses the graph-Matérn spectral density (Borovitskiy et al., 2021), the discrete analog of manifold kernels (Borovitskiy et al., 2020), as a spatial source covariance that can be combined with a temporal prior as a Kronecker product (see Appendix B).

Generative modeling of multi-channel electrophysiology. Generative modeling of electroencephalography (EEG) spans generative adversarial networks (Hartmann et al., 2018), as well as diffusion models for decoding imagined speech (Kim et al., 2023) and for synthesizing event-related potentials (Klein et al., 2024). While discriminative neural architectures build spatial structure into the network through learned spatial filters (Lawhern et al., 2018) or per-electrode embeddings (Jiang et al., 2024), generative models still start from an isotropic Gaussian. The generator, therefore, has to learn volume conduction and cross-channel covariance from the data alone. Building the source from the graph of sensor positions builds this structure into the prior, independently of the architecture and for any sensor layout.

6 Discussion

We have introduced a graph-Matérn source prior for flow-matching models of multi-channel brain signals, constructed from the sensor coordinates alone, without learned parameters. Across four flow-matching formulations and eight EEG datasets, replacing the standard isotropic Gaussian source with the graph prior reduces spectral divergence (PSD-KL) by 12% to 17% in geometric mean across datasets and up to 40% on PhysioNet-MI, while raising the correlation between real and generated phase-lag coupling (wPLI) by 0.01 to 0.03 on average. The same prior also lowers PSD-KL on whole-head MEG, intracranial EEG with patient-specific grids, and a freeway traffic-sensor network. Ablation experiments on TUAB indicate that the gain comes from the eigenvectors of a sparse local graph over sensor coordinates, rather than from variance concentration or the parametric decay of the Matérn spectrum. Priors estimated from the empirical covariance underperform isotropic noise even after shrinkage regularization.

The prior shapes only the spatial covariance and leaves all temporal structure to the velocity network. A separable spatiotemporal prior with a temporal Gaussian process (Kollovieh et al., 2025) is a natural extension. Our evaluation measures spectral match and phase coupling rather than overall sample realism, and PSD-KL compares sources within a dataset, not across datasets. Position graphs suffice for scalp and grid recordings, and manifold Matérn kernels (Borovitskiy et al., 2020) could extend the prior to folded cortical surface meshes in source-space analyses. Finally, amplitude-matched synthetic windows raise downstream balanced accuracy on Mumtaz-MDD from 60.8% to 82–83%, and whether geometry-based sources help on other clinical tasks and in continuous-time diffusion models (Rozada et al., 2026) remains to be tested.

AI use statement

In this work, we used generative AI tools for editing of the manuscript and launching experiments. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REFERENCES

Michael S. Albergo, Nicholas M. Boffi, and Eric Vanden-Eijnden. Stochastic interpolants: A unifying framework for flows and diffusions. *Journal of Machine Learning Research*, 26(209):1–80, 2025.

Anthony J. Bell and Terrence J. Sejnowski. The "independent components" of natural scenes are edge filters. *Vision Research*, 37(23):3327–3338, 1997.

Viacheslav Borovitskiy, Alexander Terenin, Peter Mostowsky, and Marc Peter Deisenroth. Matérn Gaussian processes on Riemannian manifolds. In *NeurIPS*, 2020.

Viacheslav Borovitskiy, Iskander Azangulov, Alexander Terenin, Peter Mostowsky, Marc Peter Deisenroth, and Nicolas Durrande. Matérn Gaussian processes on graphs. In *AISTATS*, 2021.

Clemens Brunner, Robert Leeb, Gernot R. Müller-Putz, Alois Schlögl, and Gert Pfurtscheller. BCI Competition 2008 – Graz Data Set A. Technical report, Graz University of Technology, 2008.

Jingjing Chen, Xiaobin Wang, Chen Huang, Xin Hu, Xinke Shen, and Dan Zhang. A large finer-grained affective computing EEG dataset. *Scientific Data*, 10(1):740, 2023.

Valentin De Bortoli, Emile Mathieu, Michael Hutchinson, James Thornton, Yee Whye Teh, and Arnaud Doucet. Riemannian score-based generative modelling. In *NeurIPS*, 2022.

Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-based generative modeling with critically-damped Langevin diffusion. In *ICLR*, 2022.

Vincent Dutordoir, Alan Saul, Zoubin Ghahramani, and Fergus Simpson. Neural diffusion processes. In *ICML*, 2023.

Theo Gasser, Petra Bächer, and Joachim Möcks. Transformations towards the normal distribution of broad band spectral parameters of the EEG. *Electroencephalography and Clinical Neurophysiology*, 53(1):119–124, 1982.

Kay Gregor Hartmann, Robin Tibor Schirrmeister, and Tonio Ball. EEG-GAN: Generative adversarial networks for electroencephalographic brain signals. *arXiv preprint arXiv:1806.01875*, 2018.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *NeurIPS*, 2017.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. In *NeurIPS Workshop on Deep Generative Models and Downstream Applications*, 2021.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *NeurIPS*, 2020.

Emiel Hoogeboom and Tim Salimans. Blurring diffusion models. In *ICLR*, 2023.

Emiel Hoogeboom, Víctor Garcia Satorras, Clement Vignac, and Max Welling. Equivariant diffusion for molecule generation in 3D. In *ICML*, 2022.

Weibang Jiang, Liming Zhao, and Bao-Liang Lu. LaBraM: Large brain model for learning generic representations with tremendous EEG data in BCI. In *ICLR*, 2024.

Gavin Kerrigan, Justin Ley, and Padhraic Smyth. Diffusion generative models in infinite dimensions. In *AISTATS*, 2023.

Agnan Kessy, Alex Lewin, and Korbinian Strimmer. Optimal whitening and decorrelation. *The American Statistician*, 72(4):309–314, 2018.

Soowon Kim, Young-Eun Lee, Seo-Hyun Lee, and Seong-Whan Lee. Diff-E: Diffusion-based learning for decoding imagined speech EEG. In *Interspeech*, 2023.

Guido Klein, Pierre Guetschel, Gianluigi Silvestri, and Michael Tangermann. Synthesizing EEG signals from event-related potential paradigms with conditional diffusion models. *arXiv preprint arXiv:2403.18486*, 2024.

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

Vernon J. Lawhern, Amelia J. Solon, Nicholas R. Waytowich, Stephen M. Gordon, Chou P. Hung, and Brent J. Lance. EEGNet: A compact convolutional neural network for EEG-based brain-computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018.

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

Sang-gil Lee, Heeseung Kim, Chaehun Shin, Xu Tan, Chang Liu, Qi Meng, Tao Qin, Wei Chen, Sungroh Yoon, and Tie-Yan Liu. PriorGrad: Improving conditional denoising diffusion models with data-dependent adaptive prior. In *ICLR*, 2022.

Yaguang Li, Rose Yu, Cyrus Shahabi, and Yan Liu. Diffusion convolutional recurrent neural network: Data-driven traffic forecasting. In *ICLR*, 2018.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *ICLR*, 2023.

Wei Liu, Jie-Lin Qiu, Wei-Long Zheng, and Bao-Liang Lu. Comparing recognition performance and robustness of multimodal deep learning models for multimodal emotion recognition. *IEEE Transactions on Cognitive and Developmental Systems*, 14(2):715–729, 2022.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *ICLR*, 2023.

Jun Ma, Banghua Yang, Wenzheng Qiu, Yunzhe Li, Shouwei Gao, and Xinxing Xia. A large EEG dataset for studying cross-session variability in motor imagery brain-computer interface. *Scientific Data*, 9(1):531, 2022.

Wajid Mumtaz. MDD patients and healthy controls EEG data (new), 2016. Figshare dataset, doi: 10.6084/m9.figshare.4244171.v2.

Guido Nolte, Ou Bai, Lewis Wheaton, Zoltan Mari, Sherry Vorbach, and Mark Hallett. Identifying true brain interaction from EEG data using the imaginary part of coherency. *Clinical Neurophysiology*, 115(10):2292–2307, 2004.

Paul L. Nunez and Ramesh Srinivasan. *Electric Fields of the Brain: The Neurophysics of EEG*. Oxford University Press, 2nd edition, 2006.

Iyad Obeid and Joseph Picone. The Temple University Hospital EEG data corpus. *Frontiers in Neuroscience*, 10:196, 2016.

Jakiw Pidstrigach, Youssef Marzouk, Sebastian Reich, and Sven Wang. Infinite-dimensional diffusion models. *Journal of Machine Learning Research*, 25(414):1–52, 2024.

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 *ICML*, 2023.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional networks for biomedical image segmentation. In *MICCAI*, 2015.

Sergio Rozada, Vimal K. B., Andrea Cavallo, Antonio G. Marques, Hadi Jamali-Rad, and Elvin Isufi. Graph-aware diffusion for signal generation. In *ICASSP*, 2026.

Gerwin Schalk, Dennis J. McFarland, Thilo Hinterberger, Niels Birbaumer, and Jonathan R. Wolpaw. BCI2000: A general-purpose brain-computer interface (BCI) system. *IEEE Transactions on Biomedical Engineering*, 51(6):1034–1043, 2004.

Neta Shaul, Ricky T. Q. Chen, Maximilian Nickel, Matt Le, and Yaron Lipman. On kinetic optimal probability paths for generative models. In *ICML*, 2023.

Fren T. Y. Smulders, Sanne ten Oever, Franc C. L. Donkers, Conny W. E. M. Quaedflieg, and Vincent van de Ven. Single-trial log transformation is optimal in frequency analysis of resting EEG alpha. *European Journal of Neuroscience*, 48(7):2585–2598, 2018.

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 (TMLR)*, 2024a.

Alexander Tong, Nikolay Malkin, Kilian Fatras, Lazar Atanackovic, Yanlei Zhang, Guillaume Huguet, Guy Wolf, and Yoshua Bengio. Simulation-free Schrödinger bridges via score and flow matching. In *AISTATS*, 2024b.

Dorien van Blooijs, Max A. van den Boom, Jaap F. van der Aar, Geertjan M. Huiskamp, Giulio Castegnaro, Matteo Demuru, Willemiek J. E. M. Zweiphenning, Pieter van Eijsden, Kai J. Miller, Frans S. S. Leijten, and Dora Hermes. Developmental trajectory of transmission speed in the human brain. *Nature Neuroscience*, 26:537–541, 2023. doi: 10.1038/s41593-023-01272-0. OpenNeuro dataset ds004080, CCEP ECoG dataset across age 4–51.

Martin Vinck, Robert Oostenveld, Marijn van Wingerden, Francesco Battaglia, and Cyriel M. A. Pennartz. An improved index of phase-synchronization for electrophysiological data in the presence of volume-conduction, noise and sample-size bias. *NeuroImage*, 55(4):1548–1565, 2011.

Vikram Voleti, Chris Pal, and Adam Oberman. Score-based denoising diffusion with non-Isotropic Gaussian noise models. *NeurIPS Workshop on Score-Based Methods*, 2022.

Daniel G. Wakeman and Richard N. Henson. A multi-subject, multi-modal human neuroimaging dataset. *Scientific Data*, 2:150001, 2015. doi: 10.1038/sdata.2015.1. OpenNeuro ds000117 (Wakeman–Henson face perception MEG/fMRI/EEG dataset).

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 *ICLR*, 2025.

Peter D. Welch. The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms. *IEEE Transactions on Audio and Electroacoustics*, 15(2):70–73, 1967.

A Experimental Details

Graph construction. Sensor coordinates are scaled so that the largest radius is one before the graph is built. The edge-weight bandwidth h^2 is one global value, the mean over all sensors of the squared distance to their k nearest-neighbors, so neither the k-NN graph nor the weights depend on the units of the coordinates (EEG head frame, MEG helmet, ECoG grid in millimeters, or Earth-centered traffic coordinates). A single bandwidth treats a dense and a sparse region of an irregular layout, such as an ECoG grid, with the same scale; a per-sensor bandwidth would adapt to local spacing, which we leave for future exploration.

Sampling. In the EEG experiments, sampling integrates the learned velocity with 50 Euler steps of the probability-flow ODE from $t{=}0$ to 1 for every method, including the $\sigma{>}0$ ones, and no score network is trained; the $\sigma{>}0$ models are therefore trained with the interpolant noise and sampled without it, identically for both sources, so the sampler is not the bridge SDE of Tong et al. (2024b). Sampling instead with $\sigma\sqrt{\Delta t}$ Gaussian noise added after each Euler step, on the runs of Appendix B, changes PSD-KL by less than 6% for every dataset-method pair of SI and SF2M on the seven datasets of Table 8 except Mumtaz with GP-SF2M (13%) and never alters the relative ranking of the two priors, indicating that conclusions are robust to the choice of sampler.

Datasets. Table 7 summarizes the eight EEG datasets of the sweep. Preprocessing, montages, and splits follow Wang et al. (2025); all recordings are resampled to 200 Hz, and each run trains on at most $20{,}000$ windows of the training split, using all available windows when fewer exist. Subject and window counts are those of the copies and splits we use. The test split holds out subjects on TUAB, TUEV, Mumtaz-MDD, BCI-IV 2a, FACED, and PhysioNet-MI; on SEED-V the copy we use covers

subjects 1–3 (seven sessions) and are split by trial within each session (trials 1–5, 6–10, and 11–15 for training, validation, and test), so every subject appears in every split; on SHU the released test split shares 30 subject-sessions with the training split, including 1,785 identical windows, so neither dataset is subject-held-out and their test-split numbers should be read as within-subject fits. The Mumtaz-MDD recordings we use cover 30 patients and 15 controls, split 31/6/8 by subject; BCI-IV 2a uses the training session of each of the nine subjects.

Table 7: The eight EEG datasets of the sweep. Windows are given in seconds and, in parentheses, in samples at 200 Hz; the native rate is the rate of the released recordings; window counts are for the splits we use before the 20,000-window training cap.

dataset	recording	subjects	channels (montage)	window	native rate	windows
TUAB (Obeid & Picone, 2016)	clinical, normal vs. abnormal	2,993 (recordings)	16 (bipolar longitudinal)	10 s (2000)	256 Hz	409,455
TUEV (Obeid & Picone, 2016)	clinical events, six classes	—	16 (bipolar longitudinal)	5 s (1000)	256 Hz	113,353
Mumtaz-MDD (Mumtaz, 2016)	depression vs. control, rest	45	19 (10–20)	5 s (1000)	256 Hz	5,087
BCI-IV 2a (Brunner et al., 2008)	motor imagery, four classes	9	22 (centro-parietal)	4 s (800)	250 Hz	2,592
FACED (Chen et al., 2023)	emotion, video clips	123	32 (10–20)	10 s (2000)	250 Hz	10,332
SHU (Ma et al., 2022)	motor imagery, two classes	52	32 (10–20)	4 s (800)	250 Hz	29,994
SEED-V (Liu et al., 2022)	emotion, five classes	3	62 (10–20)	1 s (200)	1000 Hz	17,461
PhysioNet-MI (Schalk et al., 2004)	motor imagery, four classes	109	64 (10–10)	4 s (800)	160 Hz	9,837

Hyperparameters. All runs use AdamW with learning rate 0.0002, weight decay 0.00001, gradient-norm clipping at 1.0, and batch size 64; the sweep and the ablation study keeps the learning rate constant, and only the sensitivity sweep of Appendix D and the MEG, iEEG, and PEMS-BAY runs of Appendix G use a cosine schedule. The eight-dataset sweep uses a U-Net with base channel width 48 (8.4M parameters) trained for 6k steps; TUAB's row of Table 2 and the TUAB studies of Appendices D and F use width 64 (14.9M) and 12k steps. The noise scale is σ=0.02 (0.008 on TUEV, see Appendix E), the prior uses τ=1, α=2, and k=4, the Hungarian coupling of SF2M and OT-CFM is solved exactly on each batch, and RF trains two rounds of the stated step count with one reflow on 8k generated pairs between them.

Metrics. Metrics compare 512 generated windows with 512 held-out test windows at 200 Hz (128 for MEG, iEEG, and PEMS-BAY, at the sampling rates and splits of Appendix G). In the eight-dataset sweep, each model is sampled across three generation seeds; reported values average over the five training seeds and three generation sets. For PSD-KL, log band power is computed via Welch's method (Welch, 1967) with Hann segments of $\min(256, T)$ samples and 50% overlap (200 samples on SEED-V). Because the logarithm is the transform that brings absolute band power closest to normal (Gasser et al., 1982), and for resting alpha it makes the power of single windows near-normal across windows (Smulders et al., 2018), we fit a univariate Gaussian to each channel and band, which gives the symmetric KL divergence between the real and generated fits in closed form, $\frac{1}{2}[D_{\mathrm{KL}}(\mathrm{real} \parallel \mathrm{gen}) + D_{\mathrm{KL}}(\mathrm{gen} \parallel \mathrm{real})]$, averaged across channels and bands. For wPLI, cross-channel phase synchrony is evaluated via $\mathrm{wPLI} = |\mathbb{E}[\mathrm{Im}\, S_{xy}]|/\mathbb{E}[|\mathrm{Im}\, S_{xy}|]$ (Vinck et al., 2011) using Welch cross-spectra (512-sample segments). Using only the imaginary component of coherency (Nolte et al., 2004), wPLI eliminates spurious coupling from instantaneous volume conduction. We correlate off-diagonal entries between real and generated matrices and report the band-averaged Pearson correlation (see Table 3); Gaussian white noise yields correlations within ± 0.03 across all eight datasets. We do not report the Fréchet Inception Distance because multi-channel EEG lacks a standardized feature extractor across diverse montages, and prior work found that the model with the most realistic spatial and spectral properties received the worst FID (Hartmann et al., 2018).

Downstream classifier. The EEGNet-8,2 classifier of Section 3.4 follows Lawhern et al. (2018), namely 500 epochs with validation-loss early stopping, MaxNorm 1 on the depthwise layer and 0.25 on the head, temporal kernel length $f_s/2$, Adam at learning rate 0.001, dropout 0.25, and five seeds. The generator is the sweep-configuration U-Net with a class embedding at its input, trained with classifier-free guidance (Ho & Salimans, 2021) (embedding dropped with probability 0.1) and sampled with the embedding set to the target class. To verify that the generator does not memorize training samples, we evaluate nearest-neighbor distances. After standardizing each window, the L_2 distance from a generated window to its nearest real training window has median 183 for either source, the same as between two real training windows (180), and its minimum (121 and 131 for the isotropic and graph sources) is far above the real-to-real minimum (19). The classifier is

Table 8: PSD-KL (lower is better; mean±std over five training seeds) for the isotropic, graph-Matérn, and temporal-Matérn sources at the sweep configuration. Bold marks the best source within each method. The OT-CFM and RF isotropic and graph values are those of Table 2.

	SI			SF2M			OT-CFM			RF		
dataset (ch)	isotropic	graph	temporal	isotropic	graph	temporal	isotropic	graph	temporal	isotropic	graph	temporal
TUEV (16)	3.81±1.25	3.92±1.29	**2.69±0.92**	4.32±1.34	5.28±2.99	**4.24±1.19**	5.30±2.30	**3.80±0.59**	4.04±1.16	3.19±1.2	**3.09±0.29**	3.35±1.40
Mumtaz (19)	3.55±2.09	3.58±1.05	**2.27±0.61**	4.68±2.92	4.14±1.90	**2.22±0.93**	4.97±2.06	5.51±1.66	**1.99±0.48**	6.06±3.0	3.79±1.3	**3.18±1.80**
BCI-IV 2a (22)	**3.39±0.62**	3.58±0.53	3.69±0.38	3.74±0.88	3.57±1.29	**2.91±0.48**	**3.06±0.73**	3.61±0.19	3.30±0.50	4.27±0.71	**4.24±1.3**	4.76±0.64
FACED (32)	33.34±2.87	**28.81±4.55**	32.86±5.93	32.83±3.49	31.84±7.15	**31.34±2.00**	33.60±7.20	31.50±12.00	**29.68±6.00**	32.1±3.9	**26.4±5.0**	29.30±13.33
SHU (32)	113.46±2.43	**83.17±3.71**	125.63±3.17	112.68±3.26	**83.42±4.21**	125.83±2.50	111.27±5.19	**85.36±0.71**	124.99±3.67	116±2.3	**82.9±10.3**	125.63±1.77
SEED-V (62)	32.94±0.73	**27.99±0.15**	36.80±0.80	32.85±0.75	**28.02±0.21**	36.75±0.92	32.90±0.70	**28.10±0.30**	36.74±0.82	33.3±0.4	**28.3±0.2**	36.68±0.80
PhysioNet-MI (64)	32.94±0.87	**19.57±2.26**	36.13±1.86	34.21±0.75	**20.44±3.25**	38.01±1.46	33.00±1.40	**19.90±3.90**	38.27±0.70	33.3±1.8	**22.4±1.5**	38.38±2.64

trained on real windows augmented with one generated window per training sample. Because raw generated amplitudes differ across priors, each synthetic window is scaled to match the average root-mean-square amplitude of the real training windows before concatenation.

Code and reproducibility. The training and evaluation code is released at the anonymized repository linked from the abstract, `https://anonymous.4open.science/r/graph-prior`.

B A Temporal Matérn Source Without Spatial Structure

The temporal source draws $x_0 = z\,\mathbf{V}^\top$ with $z \in \mathbb{R}^{C\times T}$ and independent entries $z_{ck} \sim \mathcal{N}(0, \psi(\omega_k))$, where $\mathbf{V}$ is the DCT-II basis of the window, $\omega_k = k/T$, and $\psi(\omega) = (1+\tau_t\omega)^{-\alpha_t}$ with τ_t=1, α_t=2, normalized to unit mean. This prior imposes temporal smoothness while leaving spatial channels uncorrelated. Combining it with the graph factor gives a separable spatiotemporal source whose covariance is the Kronecker product of the two, with samples $x_0 = \mathbf{U}\,\mathrm{diag}(\sqrt{\phi})\,z\,\mathrm{diag}(\sqrt{\psi})\,\mathbf{V}^\top$ for a standard normal matrix $z \in \mathbb{R}^{C\times T}$. With SI at the sweep configuration on FACED, this source achieves a PSD-KL of 27.8 ± 8.0 across five seeds, compared to 29.0 for the graph source alone (see Table 2). Because this difference falls within cross-seed variance and the γ band still accounts for 52% of this source's error (compared with 55% for GP-SI at σ=0.10; see Appendix H), temporal smoothing alone does not resolve the γ-band discrepancy.

Table 8 reports PSD-KL at the sweep configuration (channel width 48, 6k steps, five training seeds, test split) for SI, SF2M, OT-CFM, and RF on the seven sweep datasets. TUAB is omitted because its row in Table 2 uses the larger model configuration. The isotropic and graph baselines for SI and SF2M were retrained for this experiment and match the main benchmark within seed variance across most datasets. On TUEV, the retrained isotropic baselines reach a higher PSD-KL than in Table 2 (3.81 vs. 2.55 for SI, and 4.32 vs. 2.03 for SF2M) despite identical noise scaling (σ=0.008). An independent repetition yields SI 4.10 ± 1.07 and SF2M 3.45 ± 0.62, consistent with the training sensitivity documented in Appendix H. The relative ranking among the three priors remains identical across runs. Baseline entries for OT-CFM and RF are transcribed directly from the main sweep. RF temporal models follow the standard two-stage reflow protocol (6k steps per round, 8k reflow pairs) with the temporal source drawn in both stages. Evaluations of wPLI show that the temporal source leaves the wPLI correlation at or below the isotropic level. For SI on SEED-V, the band-averaged correlation is -0.01 for both temporal and isotropic sources, compared to 0.06 for the graph prior. On PhysioNet-MI, the temporal source yields 0.26, compared to 0.32 for isotropic noise and 0.44 for the graph prior.

C Participation Ratio vs Tikhonov Effective Dimension on TUAB

The main text uses the participation-ratio (PR) effective dimension $d_{\mathrm{eff}}(\mathbf{\Sigma}_0) = (\mathrm{tr}\,\mathbf{\Sigma}_0)^2/\,\mathrm{tr}(\mathbf{\Sigma}_0^2)$, which has no free parameter, to measure how concentrated the prior's spectrum is. On TUAB it falls from 16 for the isotropic prior to 10.1 for the graph-Matérn prior. The quantity that enters a linear-Gaussian variance bound for flow matching is instead the Tikhonov effective dimension $d_{\mathrm{eff},\lambda}(\mathbf{\Sigma}_0) = \mathrm{tr}(\mathbf{\Sigma}_0(\mathbf{\Sigma}_0 + \lambda\mathbf{I})^{-1})$, where λ is a regularization strength. Between the two priors it differs by far less at every λ, from 15.84 against 15.78 at λ=0.01 to 8.00 against 7.19 at λ=1 (see Table 9), so the PR value overstates how much the prior lowers the variance term.

Table 9: PR and Tikhonov effective dimensions of the isotropic and graph-Matérn priors on TUAB, both normalized to $\mathrm{tr}\,\mathbf{\Sigma}_0 = C$. The graph-Matérn value is smaller at every λ, but by far less than in the PR.

quantity	isotropic	graph-Matérn
PR d_{eff}	16.00	10.09
Tikhonov λ=0.01	15.84	15.78
Tikhonov λ=0.10	14.55	14.08
Tikhonov λ=0.50	10.67	9.75
Tikhonov λ=1.00	8.00	7.19
Tikhonov λ=2.00	5.33	4.82

D Graph-Prior Hyperparameter Sensitivity

Table 10 reports a 3×3 sweep of (τ, α) at fixed k=4 on TUAB with SI. Every setting except $(4, 4)$ lies in $[1.13, 1.81]$, with the best, $(0.25, 4)$, at 1.13 ± 0.08 and the default $(1, 2)$ at 1.39 ± 0.05. The sweep is a separate set of five training seeds at channel width 64 and 12k steps on the validation split with a cosine learning-rate schedule, which is why its default entry differs from the graph-Matérn references of Table 6 (1.08 ± 0.41) and Appendix F.5 (1.54 ± 0.91), both trained with a constant learning rate on their own seeds; the three values are within one another's seed variance. The configuration (τ=4, α=4) degrades performance (18.1 ± 4.8) because it puts almost all of the source variance on the lowest mode, so the source is nearly rank one and the drift does not learn to transport it to the data within the training budget. We keep $(1, 2)$ on every dataset rather than the best setting for two reasons. It is a generic setting of the graph-Matérn density of Borovitskiy et al. (2021), namely decay rate α=2 (the Matérn smoothness) and correlation length τ=1, the unit scale of the normalized Laplacian's eigenvalue range $[0, 2]$, fixed once rather than fitted by marginal likelihood as in that work. This generic setting demonstrates that the prior provides consistent gains without per-dataset tuning. In addition, the failure at (τ=4, α=4) shows that α=4, which the best setting shares, leaves no safety margin for unseen sensor layouts.

Table 10: Validation PSD-KL on TUAB with SI for the 3×3 (τ, α) sweep at k=4 (mean±std). Default $(1, 2)$ in bold.

$\tau \setminus \alpha$	1	2	4
0.25	1.62 ± 0.33	1.36 ± 0.42	1.13 ± 0.08
1.0	1.46 ± 0.37	$\mathbf{1.39 \pm 0.05}$	1.57 ± 0.15
4.0	1.44 ± 0.31	1.81 ± 0.22	18.10 ± 4.80

E Noise-Scale (σ) Calibration Across Datasets

We calibrate the noise scale through the ratio κ=$\sigma/\sigma_{\text{real}}$, where σ_{real} is the per-channel standard deviation on the training split. Setting σ=$0.08\,\sigma_{\text{real}}$ reduces per-dataset tuning to a single signal-scale estimate. A sweep of κ from 0.02 to 0.32 on TUAB with SI at the configuration of Appendix F (channel width 64, 12k steps, validation split, five seeds) finds no value between 0.04 and 0.32 distinguishable from the default 0.08 for either source, as cross-seed variance is comparable to the mean differences; only the smallest value, κ=0.02, raises the isotropic mean to 2.50. The calibration rule is therefore robust to at least a factor-of-two variation around 0.08, defining the stable interval $[0.04, 0.16]$.

Table 11 lists the σ used on each dataset together with its σ_{real} and the implied κ. In practice, σ=0.02 was used on every dataset except TUEV. FACED, whose σ_{real} is comparable to TUEV's, used σ=0.02 (κ=0.194), placing it above this interval. All other datasets fall within the interval $\kappa \in [0.04, 0.16]$, clustering closely around the target ratio of 0.08. TUEV's rescaled σ=0.008 gives κ=0.082, matching TUAB's κ=0.081 at σ=0.02.

Table 11: $\kappa = \sigma/\sigma_{\text{real}}$ across datasets at the σ used in the eight-dataset sweep. FACED's κ=0.194 lies outside the plateau $[0.04, 0.16]$ on which the rule was validated. Its per-band results are in Appendix H.

dataset	σ	σ_{real}	κ
TUAB	0.02	0.247	0.081
TUEV	0.008	0.098	0.082
Mumtaz-MDD	0.02	0.215	0.093
BCI-IV 2a	0.02	0.336	0.060
FACED	0.02	0.103	0.194
SHU	0.02	0.271	0.074
SEED-V	0.02	0.246	0.081
PhysioNet-MI	0.02	0.280	0.071

F Graph Prior Ablation Study

Unless noted otherwise, every experiment in this section trains SI on TUAB at the configuration of its row in Table 2 (dim 64, 12k steps) and reports PSD-KL on the validation split. The isotropic (1.85 ± 0.18) and graph-Matérn (1.08 ± 0.41) reference values therefore differ from the test-split values in that table (1.43 and 1.16). We use SI because its independent coupling adds no alignment step between the source and the drift that could mask a change of the source.

F.1 Replacing Only the Eigenvectors: Shuffled Positions and Random Bases

Rotating the 3-D sensor coordinates does not change the graph prior, since rigid rotations preserve pairwise Euclidean distances, the k-NN graph, and its eigenvectors U in channel space. To isolate the eigenvectors from the spectrum, the three priors in this group retain the Matérn density $\phi(\lambda) = (1 + \tau\lambda)^{-\alpha}$ and normalization unchanged (d_{eff}=10.1), altering only the spatial directions.

Shuffled sensor positions. Permuting sensor coordinates before constructing the k-NN graph attaches the same graph topology and spectrum to incorrect channels. Across 25 random permutations, the mean PSD-KL is 3.70 ± 2.79 (median 2.71). 64% of the permutations perform worse than the uninformative isotropic source (1.85 ± 0.18), 84% are worse than the true position graph (1.08 ± 0.41), and the spread across permutations is nearly seven times larger (2.79 vs. 0.41). The wide spread plausibly reflects how much local neighborhood structure each permutation happens to preserve.

Uniformly random orthogonal basis. Replacing U with a uniformly random orthogonal matrix $Q \in \mathbb{R}^{C \times C}$ preserves the exact Matérn eigenvalues but randomizes channel axes, yielding 4.42 ± 2.97 and confirming that concentrating variance along arbitrary axes harms generation.

Empirical covariance eigenvectors. Pairing the eigenvectors of the empirical data covariance $\hat{\Sigma}_{\text{emp}}$ with the smooth graph-Matérn spectrum yields 2.76 ± 2.93. Even with a smoothly decaying spectrum and no small eigenvalues, data-driven PCA modes fail to substitute for the physical sensor eigenbasis.

F.2 Concentrating Variance in the Wrong Directions: Empirical Covariance and Shrinkage

The empirical-covariance prior tests whether a concentrated source spectrum suffices without montage geometry. We estimate $\hat{\Sigma}_{\text{emp}}$ over all time samples of the TUAB training windows, eigendecompose it as $\hat{\Sigma}_{\text{emp}} = U_{\text{emp}} \Lambda_{\text{emp}} U_{\text{emp}}^{\top}$, scale Λ_{emp} so that its mean eigenvalue is one as for the graph prior, and draw $x_0 = U_{\text{emp}} \operatorname{diag}(\sqrt{\lambda_{\text{emp}}})\, z$. This prior uses no montage geometry and is more concentrated than the graph-Matérn prior (d_{eff}=5.2 against 10.1, with two eigenvalues below 0.001). If concentration alone explained the gain, this prior would outperform the graph-Matérn prior. Instead, its PSD-KL is 4.61 ± 2.40, worse than both the isotropic source (1.85 ± 0.18) and all priors that

replace only the eigenvectors (see Table 6). The gain therefore requires placing variance along the physical sensor layout rather than simply concentrating variance.

Regularization via Ledoit–Wolf shrinkage. Because $\hat{\Sigma}_{\text{emp}}$ is estimated from finite training data, we test whether ill-conditioned or noisy sample eigenvalues cause this failure. We regularize the covariance as $\hat{\Sigma}_{\text{LW}} = (1-\rho)\hat{\Sigma}_{\text{emp}} + \rho\bar{\lambda}\,\mathbf{I}$, where $\bar{\lambda} = \text{tr}(\hat{\Sigma}_{\text{emp}})/C$ and ρ is the optimal Ledoit–Wolf shrinkage intensity (Ledoit & Wolf, 2004). Shrinking the empirical covariance flattens its spectrum and raises the effective dimension from 5.2 to 6.6, but preserves its eigenvectors. At 3.56 ± 2.43, the regularized prior remains worse than isotropic noise (1.85 ± 0.18), confirming that the failure lies in the data-driven eigenvectors rather than eigenvalue noise.

Synthetic Gaussian target. To test whether covariance mismatch between a Gaussian source and the target explains the failure, we replace the EEG target with a synthetic Gaussian target, white in time, matching the empirical TUAB channel covariance. On this target, all six tested Gaussian sources reach a similar PSD-KL (0.0020–0.0032; unit isotropic 0.0020 ± 0.0004, empirical covariance 0.0022 ± 0.0005, Ledoit–Wolf 0.0022 ± 0.0004). The velocity field absorbs any mismatch between Gaussian distributions, so the empirical prior's failure on real EEG likely comes from temporal dependencies and non-Gaussian structure that this target lacks.

F.3 Spectrum Shape: Smoothness and Mode Coverage

We next evaluate the functional form of the spectral roll-off while holding the eigenvectors of the position graph fixed.

Heat-kernel spectrum. Replacing the polynomial Matérn density with an exponential heat kernel $\phi(\lambda) = \exp(-\tau\lambda)$ on the same Laplacian yields 1.38 ± 0.86. This retains most of the benefit of the graph-Matérn prior (1.08 ± 0.41), although placing more variance on higher-frequency modes (d_{eff}=12.0 vs. 10.1) raises the standard deviation over seeds from 0.41 to 0.86. Appendix D provides a sensitivity analysis over the Matérn hyperparameters (τ, α).

Hard low-pass spectrum. Assigning equal variance to the $C/2 = 8$ smoothest eigenmodes and setting the remaining upper half strictly to zero (d_{eff}=8.0) fails (71.4 ± 12.5). Because flow matching learns a smooth invertible map, it cannot map a source with zero variance on a subspace to a target that has variance in every direction.

F.4 Graph Sparsity and Construction: Fully Connected vs. Local Graphs

We examine how graph connectivity and construction influence generation.

Fully connected Gaussian graph. Connecting every sensor pair with a dense Gaussian weight $\exp(-\|p_i - p_j\|^2/h^2)$ rather than restricting edges to k-NN yields 2.53 ± 1.81 (d_{eff}=11.2), performing worse than isotropic noise. Geometric distance alone is therefore insufficient; the inductive bias requires the eigenvectors of a sparse, local graph.

Correlation graph in place of coordinates. Replacing coordinate distances with training-set cross-channel correlations $|\text{corr}(x_i, x_j)|$ in a k=4 nearest-neighbor graph yields 1.39 ± 1.70 across ten seeds (median 0.90, d_{eff}=11.0). The correlation graph shares only 36% of its edges with the spatial coordinate graph; its strongest edges link distant cross-scalp channels (e.g., frontal–occipital pairs reaching $|\text{corr}| \approx 0.99$), likely due to broad-band artifacts shared across the montage rather than local coupling. A graph built from sensor positions has no such long-range edges.

F.5 Whitening and a Graph Convolution in the Drift

Two baselines place spatial structure in the data or in the drift instead of in the source. The first whitens every window with a fixed zero-phase (ZCA) whitening map $\hat{\Sigma}^{-1/2}$ (Bell & Sejnowski, 1997; Kessy et al., 2018), the symmetric inverse square root of the channel covariance estimated from 1024 training windows, and trains with the isotropic source on the whitened data, undoing the

Table 12: PSD-KL on the TUAB validation split (lower is better; mean±std over five training seeds) for the whitening and graph-mixer baselines with SI at dim 64, 12k steps. Bold denotes the lowest mean; every difference is within the spread over seeds.

configuration	PSD-KL
isotropic source, plain drift	1.23 ± 0.36
ZCA-whitened data, isotropic source	$\mathbf{0.98 \pm 0.38}$
graph-Matérn source, plain drift	1.54 ± 0.91
isotropic source, graph-mixer drift	1.46 ± 0.54
graph-Matérn source, graph-mixer drift	1.67 ± 1.23

Table 13: Per-patient PSD-KL on the CCEP ECoG dataset (mean±std over five training seeds; lower is better). The channel counts, the number of ECoG channels with valid 3-D coordinates, are 104 (UMCU01), 88 (UMCU02), 72 (UMCU03), 112 (UMCU04), and 96 (UMCU05). Bold denotes the better source within each method for each patient and for the mean.

	σ>0 Hungarian		σ>0 independent		σ=0 Hungarian		σ=0 indep+reflow	
patient	SF2M	GP-SF2M	SI	GP-SI	OT-CFM	GP-OT-CFM	RF	GP-RF
UMCU01	58.0±0.6	**35.5±1.8**	57.5±2.2	**35.9±1.7**	56.2±1.5	**34.9±1.9**	67.4±0.8	**48.6±0.9**
UMCU02	6.0±0.2	**3.4±0.1**	5.9±0.3	**3.6±0.3**	5.8±0.2	**3.6±0.1**	9.0±0.9	**5.0±0.4**
UMCU03	2.5±0.1	**1.6±0.1**	2.5±0.1	**1.7±0.1**	2.5±0.1	**1.6±0.1**	3.7±0.2	**2.3±0.1**
UMCU04	76.4±1.1	**47.6±1.6**	75.8±2.0	**44.8±2.5**	74.9±1.7	**47.3±3.1**	89.9±1.4	**58.2±3.5**
UMCU05	13.5±0.3	**7.8±0.2**	14.2±0.9	**7.5±0.3**	13.0±0.7	**7.3±0.4**	18.0±0.5	**10.5±0.4**
mean across patients	31.3	**19.2**	31.2	**18.7**	30.5	**18.9**	37.6	**24.9**

map before evaluation. The second keeps the drift's 1D U-Net and inserts a channel mixer $(\mathbf{I} - \beta\mathbf{L})$ with a learned scalar β at its input and output, with $\mathbf{L}$ the same k-NN Laplacian the prior uses, under either source. All five configurations of Table 12 are trained with SI on TUAB in one run per seed, so they share the same training batches, evaluation set, and isotropic reference, with the bipolar-pair coordinates, the validation split, batch size 64, and five seeds at the section's configuration (dim 64, 12k steps). The five configurations are within the spread over seeds of one another, isotropic 1.23 ± 0.36, whitened 0.98 ± 0.38, graph-Matérn 1.54 ± 0.91, and 1.46 ± 0.54 and 1.67 ± 1.23 with the mixer, so at this budget placing spatial structure in the data or in the drift does not separate from the isotropic reference on TUAB. The isotropic reference of this run is itself below the section's (1.85 ± 0.18), which comes from a separate set of training seeds.

G Details of the Other Modalities

The three modalities of Section 3.3 use the protocol of Section 3.1 with the graph and the training setup of each modality given below. Table 13 gives the per-patient values behind the iEEG row of Table 4.

Whole-head MEG. We use the Wakeman–Henson face-perception cohort (Wakeman & Henson, 2015), 16 subjects recorded with an Elekta Vectorview helmet (102 magnetometers, one graph shared across subjects), resampled to 250 Hz and segmented into 4 s windows. One model per method, source, and seed is trained on all 16 subjects for 48k steps and evaluated on the last 20% of each subject's recording. Unlike the EEG runs, SF2M and SI on MEG are sampled with $\sigma\sqrt{\Delta t}$ Gaussian noise added after each Euler step. All three modalities train a three-level U-Net of channel width 48 (channel multipliers 1, 2, 4) with batch size 32, σ=0.02, and a cosine learning-rate schedule, and evaluate 128 generated against 128 held-out windows. The step count is chosen where PSD-KL improvements stabilize, as extending training from 12k to 48k steps shifts relative gains by at most 0.3 percentage points across methods.

Intracranial EEG. We use five CCEP ECoG (van Blooijs et al., 2023) patients (UMCU01–05) whose ECoG grids have 72 to 112 channels (see Table 13). Each patient has a unique grid and channel

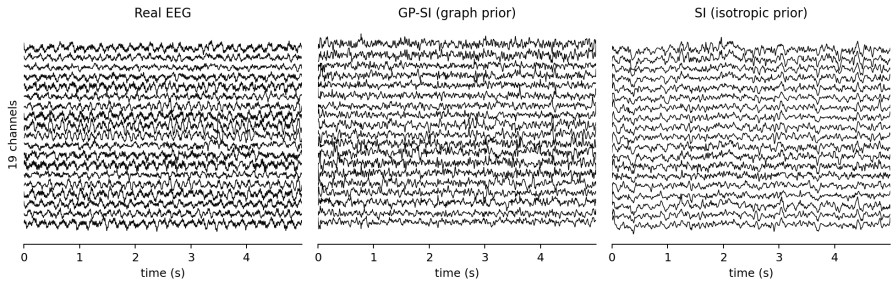

Figure 3: **Qualitative effect of the graph prior on Mumtaz-MDD** (19 channels). Real EEG (left), GP-SI (middle), and SI (right). GP-SI reproduces the slow-wave dynamics and channel coupling of real signals, whereas isotropic SI produces decoupled channel noise.

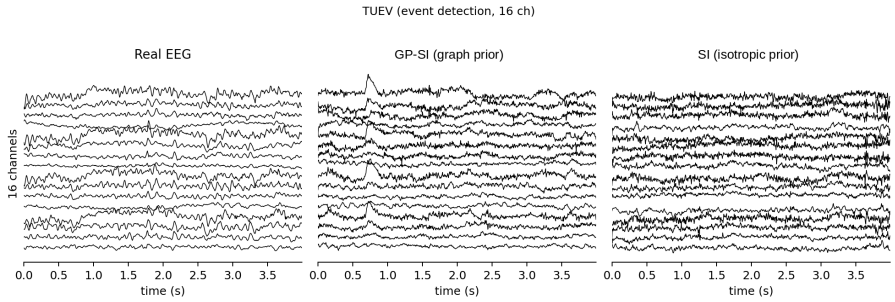

Figure 4: **Qualitative effect of the graph prior on TUEV** (16 channels). Real EEG (left), GP-SI (middle), and SI (right). GP-SI reproduces the slow-wave dynamics and channel coupling of real signals, whereas isotropic SI produces decoupled channel noise.

count, so a shared Laplacian is not defined. We train one model per patient, method, source, and seed on that patient's recordings (band-pass 0.5–100 Hz, 50 Hz notch, 250 Hz sampling rate, 8 s windows) using the graph Laplacian constructed from that patient's electrode coordinates. Each model trains for 6k steps; RF trains two rounds of 6k steps with one reflow on $2{,}048$ generated pairs.

PEMS-BAY traffic. The sensor network of Li et al. (2018) has 325 freeway loop sensors in the San Francisco Bay Area. We convert latitude and longitude to Earth-centered coordinates for the k-NN graph and use windows of $T{=}256$ samples. PEMS-BAY is sampled every 5 minutes, so its Nyquist frequency of 1.67 mHz lies far below the EEG bands, and PSD-KL instead uses five uniform bands of width 0.33 mHz. One model per method, source, and seed trains for 6k steps; RF trains two rounds of 6k steps with one reflow on $4{,}096$ generated pairs. The symmetrized $k{=}4$ graph is connected (its Laplacian has a single zero eigenvalue) on the 325-sensor network and on every ECoG grid, which we verify for each layout before training.

H Qualitative and Per-Band Details of the Eight-Dataset Sweep

The eight-dataset sweep of Table 2 trains the four methods with the isotropic and the graph source on each dataset.

Per-dataset qualitative waveforms. Figures 3 and 4 repeat the qualitative comparison of Figure 2 on Mumtaz and TUEV. On Mumtaz (depression, 19 channels) the graph prior retains the slow waves and inter-channel coherence of real EEG, whereas the isotropic prior produces noisier channels that are only weakly coupled. On TUEV, the graph prior preserves the slow-wave amplitude variability of real EEG and inter-channel coupling, while the isotropic prior produces unstructured noise with decoupled channels.

Table 14: Per-band PSD-KL on SHU and SEED-V for all four methods (test split; the $\sigma>0$ methods at the default $\sigma=0.02$); bold marks the best value per band per dataset.

dataset	variant	coupling	δ	θ	α	β	γ	avg
SEED-V (62ch)	SF2M	Hung, $\sigma>0$	1.27	7.75	14.18	56.20	84.82	32.84
	GP-SF2M	Hung, $\sigma>0$	1.25	6.41	11.66	47.20	73.73	28.05
	SI	indep, $\sigma>0$	1.31	7.78	14.20	56.36	85.01	32.93
	GP-SI	indep, $\sigma>0$	**1.22**	**6.40**	**11.65**	**47.14**	**73.66**	**28.01**
	OT-CFM	Hung, $\sigma=0$	1.28	7.79	14.22	56.29	84.90	32.90
	GP-OT-CFM	Hung, $\sigma=0$	1.26	6.41	11.67	47.22	73.76	28.06
	RF	indep+reflow, $\sigma=0$	1.32	7.93	14.47	57.14	85.63	33.30
	GP-RF	indep+reflow, $\sigma=0$	1.26	6.53	11.85	47.76	75.08	28.50
SHU (32ch)	SF2M	Hung, $\sigma>0$	6.04	24.38	28.87	168.65	337.45	113.08
	GP-SF2M	Hung, $\sigma>0$	3.57	16.67	18.71	**111.25**	256.59	**81.36**
	SI	indep, $\sigma>0$	6.10	24.59	29.01	171.71	340.40	114.36
	GP-SI	indep, $\sigma>0$	**3.31**	**16.27**	**18.39**	114.40	261.43	82.76
	OT-CFM	Hung, $\sigma=0$	5.83	23.81	28.04	162.59	336.05	111.27
	GP-OT-CFM	Hung, $\sigma=0$	3.41	16.53	18.99	122.27	265.61	85.36
	RF	indep+reflow, $\sigma=0$	6.47	25.13	29.76	173.86	337.45	114.53
	GP-RF	indep+reflow, $\sigma=0$	3.61	16.43	19.31	116.05	**254.71**	82.02

Per-band breakdown on SHU and SEED-V. Table 14 reports per-band PSD-KL on SHU (32 channels) and SEED-V (62 channels) for all four methods, with the $\sigma>0$ methods at the default $\sigma=0.02$. The SF2M, SI, and OT-CFM rows are computed from the per-seed band values of the benchmark runs in Table 2 (five training seeds, three generated sets each), so their avg column matches that table. The RF rows come from a separate five-seed run of the same reflow protocol whose per-band values were stored; their avg column (114.53 and 82.02 on SHU, 33.30 and 28.50 on SEED-V) sits within the cross-seed variance of the RF and GP-RF values in Table 2 (116 ± 2.3 and 82.9 ± 10.3; 33.3 ± 0.4 and 28.3 ± 0.2). Every method with the isotropic source yields comparable averages regardless of coupling (32.8 to 33.3 on SEED-V and 111.3 to 114.5 on SHU), and every method with the graph source yields consistently lower error (28.0 to 28.5 on SEED-V and 81.4 to 85.4 on SHU). The source prior matters more than the choice of coupling, since the four graph-source averages lie within 2% of each other on SEED-V and within 5% on SHU, while the gap between the two sources is 14% to 15% on SEED-V and 23% to 28% on SHU. On SHU, the graph source reduces PSD-KL across all frequency bands for every method, with the γ term dropping from 340 to 261 for SI.

Per-band breakdown on TUEV and FACED. TUEV is the only dataset on which the graph prior makes SF2M worse ($2.03 \rightarrow 3.85$ at the sweep configuration; see Table 2). The increase is spread over all five bands (from $+82\%$ in δ and α to $+241\%$ in β, five seeds at $\sigma=0.008$), and the cross-seed variance of GP-SF2M is comparable to the shift, indicating training variance rather than a localized spectral failure. On FACED the prior lowers PSD-KL for all four methods at $\sigma=0.02$ (see Table 2). At $\sigma=0.10$, where high noise exacerbates spectral divergence, GP-SI matches or outperforms isotropic baselines in δ, θ, and α, but diverges in β and γ (see Figure 5); γ alone contributes more than half of the five-band average. A separable spatiotemporal source on FACED is reported in Appendix B.

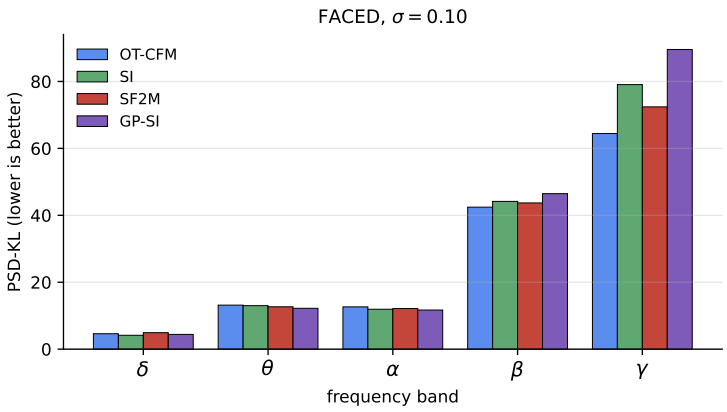

Figure 5: Per-band PSD-KL on FACED for OT-CFM (σ=0) and for SI, SF2M, and GP-SI at σ=0.10 (mean over seeds; lower is better).