A FLOW MATCHING FRAMEWORK FOR NEURAL REPRESENTATIONAL DISSIMILARITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Neural representational dissimilarity quantifies differences between neural response distributions, and is essential for comparing neural codes across stimuli, brain areas, tasks, and models. Commonly used distance metrics involve different assumptions and are estimated with separate methods. Here, we show that a variety of distance metrics can be unified under a flow matching framework developed in deep generative models. That is, these distances arise as Jeffreys divergences under different velocity constraints. We find that flow matching has advantages for estimating distances involving complicated distributions and continuous variables. Furthermore, this framework enables the design of new distance metrics in a principled way. Together, flow matching provides a unified approach for understanding, estimating, and designing neural representational dissimilarity metrics.

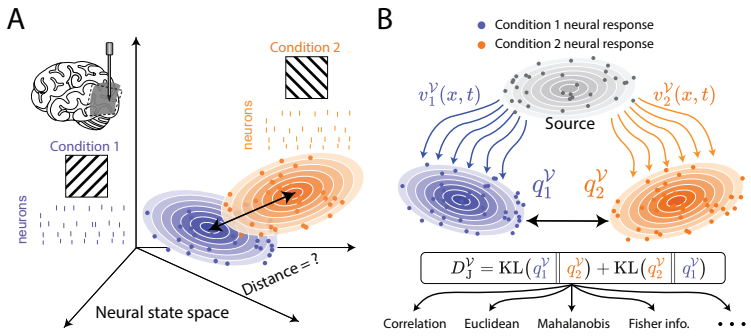

Figure 1: **Overview of the proposed flow matching framework for neural representational dissimilarity.** (**A**) Different stimuli evoke condition-specific neural responses, producing different population-response distributions whose dissimilarity is to be quantified. (**B**) Flow matching learns velocity fields $v_1^{\mathcal{V}}(x,t)$ and $v_2^{\mathcal{V}}(x,t)$ that transport a shared source distribution to the fitted distributions $q_1^{\mathcal{V}}$ and $q_2^{\mathcal{V}}$. Here, we define representational dissimilarity as the Jeffreys divergence between $q_1^{\mathcal{V}}$ and $q_2^{\mathcal{V}}$. We prove that different choices of $\mathcal{V}$ correspond to different distance metrics.

1 INTRODUCTION

A central question in neuroscience is how the brain represents information through stochastic neural population responses. Representing information means not only "storing" it but also "structuring" it in an appropriate format to support downstream computation given the biological constraints (Kriegeskorte & Kievit, 2013; Kriegeskorte & Wei, 2021). How can we study representational structures? Representational similarity analysis (RSA) takes a geometric perspective, probing representational structure by measuring pairwise distances between neural population response distributions across conditions, such as stimuli or task states (Kriegeskorte et al., 2008a; Nili et al., 2014). RSA has been widely used to study how neural representations support perception and task performance (Kriegeskorte et al., 2008a; Cichy et al., 2014; Diedrichsen & Kriegeskorte, 2017) and to compare representational systems across species and computational models, including deep neural network models (Kriegeskorte et al., 2008b; Kriegeskorte, 2009; Kornblith et al., 2019).

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.

A key step in RSA is to define and estimate representational distances. Because neural responses are stochastic, this often becomes a mathematical question of how to compare distributions. No single distance metric is best for every purpose as different metrics reveal different properties of the distributions (Walther et al., 2016; Diedrichsen & Kriegeskorte, 2017; Kriegeskorte & Wei, 2021). For example, correlation distance compares the relative pattern of mean activity across neurons. Euclidean distance measures separation between mean response vectors. Mahalanobis distance additionally accounts for response covariance. Kullback-Leibler (KL) divergence compares full response distributions. Lastly, Fisher information measures how sensitive the response distribution changes with a small change of the stimulus variable (Kullback & Leibler, 1951; Amari, 2016).

Given the diversity of these useful distance metrics, a natural question arises: can they be unified under a single mathematical framework? Such a framework, if exists, could offer several benefits. First, it would make the statistical assumptions underlying different metrics more explicit, clarifying their interpretation. Second, it would provide a principled way to design new distance metrics. Finally, it could provide a unified estimator for these metrics, so that improvements to the estimator could benefit multiple metrics rather than requiring separate efforts for each one.

Here, we provide such a framework using flow matching. Flow matching is a modern deep generative modeling method that is widely used in image, video, and language generation (Lipman et al., 2023; Esser et al.; Davtyan et al., 2023; Hu et al., 2024). It starts by generating a data sample from a simple source distribution (e.g., a Gaussian distribution) and then evolves the sample toward a target distribution through a velocity field. In our flow-matching framework for neural dissimilarity, we define the representational distance as the Jeffreys divergence between two target neural-response distributions under a given velocity constraint. We mathematically prove that this Jeffreys divergence reduces to existing distance metrics under different velocity-field constraints (Fig. 1). We show that flow matching offers advantages when estimating distances involving complex distributions and continuous variables, and allows the principled design of new distance metrics.

2 Background and Related Work

Representational dissimilarity analysis. RSA summarizes neural, behavioral, or model representations by a representational dissimilarity matrix (RDM), whose entries measure the pairwise separation between conditions (Kriegeskorte et al., 2008a; Kriegeskorte, 2009; Kriegeskorte & Kievit, 2013; Nili et al., 2014). Common RDM entries include correlation, Euclidean, Mahalanobis, and cross-validated Mahalanobis distances; these choices differ in how they treat response scale, noise covariance, and finite-sample bias (Walther et al., 2016; Diedrichsen & Kriegeskorte, 2017).

Fisher information estimation. Fisher information measures how rapidly $p_\theta(x)$ changes with a continuous stimulus or behavioral variable, θ (Fisher, 1922; Amari, 2016; Seung & Sompolinsky, 1993; Pouget et al., 2000). Its inverse lower-bounds the variance of unbiased estimators, making Fisher information a local measure of discriminability (Seriès et al., 2009; Rao, 1945; Brunel & Nadal, 1998; Abbott & Dayan, 1999; Averbeck et al., 2006). With repeated trials at nearby stimulus values, Fisher information can be estimated from tuning-curve derivatives and response covariance or from decoder-based estimator (Kanitscheider et al., 2015; Kohn et al., 2016). Continuous and naturalistic experiments, however, have no exact repeats. Estimation procedures need to exploit information shared across nearby values of the condition θ, for example by using Gaussian processes (Rasmussen & Williams, 2006; Ye & Wessel, 2025; Nejatbakhsh et al., 2023). However, Gaussian processes scale poorly (Bruinsma et al., 2020). Our flow-matching framework unifies Fisher information estimation with the other distance metrics that have traditionally been treated separately.

Continuous normalizing flows and flow matching. Normalizing flows model densities through invertible transformations (Rezende & Mohamed, 2015; Papamakarios et al., 2021), while continuous normalizing flows transport samples through an ODE (Chen et al., 2018). Roughly speaking, a sample begins at $x_0 \sim \rho_0$, follows the velocity field v over time, and arrives at a distribution intended to approximate the observed data. In the original continuous-flow formulation, the velocity parameters are trained by maximum likelihood, which requires solving the ODE and estimating the divergence of the velocity field during optimization, which can be slow (Chen et al., 2018; Grathwohl et al., 2019). Flow matching reformulates this training problem. Instead of maximizing likelihoods, one specifies a probability path between source samples and data samples, computes the target velocity

along that path, and trains v by regressing onto those velocities (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023). Flow matching thus provides a practical way to learn a transport map from a source distribution to an empirical response distribution.

3 A FLOW MATCHING FRAMEWORK FOR COMPUTING REPRESENTATIONAL DISTANCE

Flow matching learns a time-dependent velocity field that transports samples from a simple *source* distribution to a *target* distribution (Lipman et al., 2023). Consequently, different constraints on the velocity field yield different approximations to the target distribution. Let $x \in \mathbb{R}^d$ denote a neural population response vector, where d is the number of neurons, and let $p_c(x)$ denote the unknown target distribution under condition $c \in \mathcal{C}$. Starting from a source sample $X_0 \sim \rho_0$, sampling follows the ODE

$$\frac{dX_t}{dt} = v_c^{\mathcal{V}}(X_t, t), \qquad t \in [0, 1], \tag{1}$$

where the velocity field $v_c^{\mathcal{V}}(x, t)$ specifies the direction and speed of the sample at each flow time and $\mathcal{V}$ denotes the chosen velocity class. Integrating Eq. (1) from $t = 0$ to $t = 1$ defines a transformation $T_{c,1}^{\mathcal{V}}$ and the fitted distribution

$$q_c^{\mathcal{V}} = (T_{c,1}^{\mathcal{V}})_{\#}\rho_0, \tag{2}$$

so that $T_{c,1}^{\mathcal{V}}(X_0) \sim q_c^{\mathcal{V}}$ when $X_0 \sim \rho_0$. The fitted distribution $q_c^{\mathcal{V}}$ approximates p_c. Here, t is an artificial flow-matching coordinate, not physical time.

In general, our framework has three steps: (1) choose a source distribution and a velocity class $\mathcal{V}$; (2) learn the velocity field by flow matching to approximate the target distributions; and (3) use the learned model to compute the Jeffreys divergence between the approximated distributions as the representational distance.

Notation. We use $a, b \in \mathcal{C}$ for a pair of categorical conditions being compared and $\mu_c = \mathbb{E}_{X \sim p_c}[X]$ for the mean response under condition c. We use $\theta \in \Theta$ for a continuously varying condition, such as stimulus orientation, physical time, or behavior. For simplicity, we describe the framework below using the categorical condition c; the same framework applies to the continuous condition θ.

1. Choosing a source distribution and velocity class. We first choose a source distribution ρ_0 and a velocity class $\mathcal{V}$. The source specifies the initial distribution, which is typically shared across conditions. Meanwhile, the velocity class specifies the transformations allowed from the source to each target. For example, an unconstrained velocity field, parameterized by a neural network, permits flexible nonlinear transformations, whereas a constant, state-independent velocity restricts the transformation to a translation. These choices determine which properties of the target distributions can be represented (see Fig. 2) and, consequently, which representational distance is induced. Section 4 presents the specific choices used to recover different distance metrics.

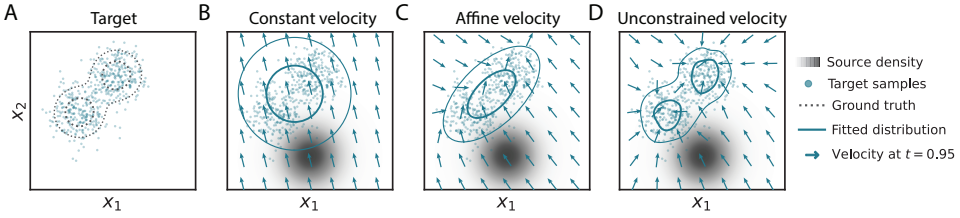

Figure 2: **Velocity classes $\mathcal{V}$ used in flow matching determine the fitted distributions** $q_c^{\mathcal{V}}$ (also see Table 1). (A) A two-component Gaussian-mixture target, shown as samples and dotted density contours. (B) A constant velocity translates the Gaussian source (black region) without changing its shape. (C) An affine velocity allows an affine transformation of the source distribution, yielding a Gaussian approximation to the target. (D) An unconstrained velocity captures the two modes of the target. Arrows indicate velocity direction, not magnitude.

2. Flow-matching training. Given the source distribution and velocity class, flow matching learns the velocity field with a simple regression objective (Lipman et al., 2023; Albergo & Vanden-Eijnden, 2023). During training, we draw source samples $X_0 \sim \rho_0$ and observed responses $X_1^{(c)} \sim p_c$ and connect them along a predefined path

$$X_t^{(c)} = \alpha_t X_0 + \beta_t X_1^{(c)}, \qquad (\alpha_0, \beta_0) = (1, 0), \quad (\alpha_1, \beta_1) = (0, 1), \tag{3}$$

where α_t and β_t are prescribed interpolation schedules controlling the contributions of the source and target samples at flow time t. The boundary conditions ensure that the path begins at the source sample and ends at the recorded response. Its target instantaneous velocity is

$$U_t^{(c)} = \dot{\alpha}_t X_0 + \dot{\beta}_t X_1^{(c)}, \tag{4}$$

where the dots denote derivatives with respect to t. A neural network receives $(X_t^{(c)}, t, c)$ and is trained to predict this target velocity. The flow-matching objective is

$$\mathcal{L}_{\mathrm{FM}}(v) = \mathbb{E}_{\substack{C\sim\pi_{\mathcal{C}},\ t\sim\mathrm{Unif}[0,1] \\ X_0\sim\rho_0,\ X_1^{(C)}\sim p_C}} \left[\left\| v_C(X_t^{(C)}, t) - U_t^{(C)} \right\|^2 \right], \tag{5}$$

where $C \sim \pi_{\mathcal{C}}$ denotes a randomly sampled categorical condition. For a continuous condition, we instead sample $\Theta \sim \pi_\Theta$ and replace C, p_C, $X_1^{(C)}$, and v_C by Θ, p_Θ, $X_1^{(\Theta)}$, and v_Θ, respectively. The population minimizer $v^{\mathcal{V}} = \arg\min_{v\in\mathcal{V}} \mathcal{L}_{\mathrm{FM}}(v)$ is learned by backpropagation through the velocity network.

Optionally, the learned velocity field can be fine-tuned by maximizing the log likelihood (Section B.1). We use this fine-tuning only for the new distance metric introduced in Supplementary Section C, where it is practically useful for optimization; all other results use flow-matching training alone for simplicity.

3. Jeffreys-divergence computation. We define the representational distance between conditions a and b as the Jeffreys divergence between their *fitted* distributions (Kullback & Leibler, 1951; Amari, 2016),

$$D_{\mathcal{V}}(a, b) = D_{\mathrm{KL}}\big(q_a^{\mathcal{V}} \| q_b^{\mathcal{V}}\big) + D_{\mathrm{KL}}\big(q_b^{\mathcal{V}} \| q_a^{\mathcal{V}}\big), \tag{6}$$

where each directed KL divergence is

$$D_{\mathrm{KL}}\big(q_a^{\mathcal{V}} \| q_b^{\mathcal{V}}\big) = \mathbb{E}_{x\sim q_a^{\mathcal{V}}} \left[\log q_a^{\mathcal{V}}(x) - \log q_b^{\mathcal{V}}(x)\right]. \tag{7}$$

When the source distribution is Gaussian and the velocity class $\mathcal{V}$ is affine (linear in the state), the fitted distributions $q_c^{\mathcal{V}}$ can be obtained by solving deterministic ODEs for their means and covariances (Supplementary Section A.3), from which the Jeffreys divergence can be computed.

In other cases, we estimate the Jeffreys divergence using Monte Carlo. We draw samples $X_i^{(a)} \sim q_a^{\mathcal{V}}$ and $X_i^{(b)} \sim q_b^{\mathcal{V}}$ for $i = 1, \ldots, M$ by integrating the sampling ODE in Eq. (1) and compute

$$\widehat{D}_{\mathcal{V}}(a, b) = \frac{1}{M}\sum_{i=1}^{M}\big[\log q_a^{\mathcal{V}}(X_i^{(a)}) - \log q_b^{\mathcal{V}}(X_i^{(a)})\big] + \frac{1}{M}\sum_{i=1}^{M}\big[\log q_b^{\mathcal{V}}(X_i^{(b)}) - \log q_a^{\mathcal{V}}(X_i^{(b)})\big]. \tag{8}$$

These log densities are computed using the continuous change-of-variables identity in Eq. (111).

4 Flow matching with different velocity constraints lead to different representational metrics

We now present our main theoretical results, *i.e., various commonly used representational distance metrics arise from our flow matching framework as Jeffreys divergences under different velocity classes.* These correspondences are summarized in Table 1. For each velocity class, the proof follows a similar general strategy. We substitute the analytical form of the constrained velocity field into the population flow-matching loss and solve for its minimizer, which determines the induced approximate distribution. Evaluating the Jeffreys divergence between the resulting condition-specific

Table 1: Distance metrics and their corresponding velocity classes. We use c for a categorical condition and θ for a continuous condition. All rows use the shared standard Gaussian source $\rho_0 = \mathcal{N}(0, I)$ except the starred geometry-template distance, which uses a source distribution concentrated around a low-dimensional manifold (Supplementary Section C). The cosine and correlation rows equal $2r^2$ times their conventional definitions. In the Mahalanobis row, A_t is shared across conditions. In the linear-Fisher row, $\bar{A}_{\theta,t}$ is shared locally across nearby conditions (see Supplementary Section A.4).

Velocity field	Induced distance
$v_c(x,t) = \dot{\beta}_t b_c$	Euclidean distance
$v_c(x,t) = \dot{\beta}_t b_c,\ \|b_c\| = r$	Cosine distance
$v_c(x,t) = \dot{\beta}_t b_c,\ \mathbf{1}^\top b_c = 0,\ \|b_c\| = r$	Correlation distance
$v_c(x,t) = \dot{\beta}_t b_c + A_t(x - \beta_t b_c)$	Mahalanobis distance
$v_c(x,t) = \dot{\beta}_t b_c + A_{c,t}(x - \beta_t b_c)$	Gaussian Jeffreys divergence
$v_\theta(x,t) = \dot{\beta}_t b_\theta + \bar{A}_{\theta,t}(x - \beta_t b_\theta)$	Linear Fisher information
$v_c(x,t),\ v_\theta(x,t) \in \mathcal{V}_{\text{unconstrained}}$	Jeffreys divergence; full Fisher information
$v_c(x,t;\tau),\ v_\theta(x,t;\tau)$	Auxiliary τ-dependent distance
$v_c(x,t) = a_{c,t} + \Omega_{c,t}(x - o_{c,t}) + \lambda_{c,t}(x - o_{c,t}),\ {\Omega_{c,t}}^\top = -\Omega_{c,t}$	Geometry-template Jeffreys divergence*

*: new distance introduced in this paper; dark red: outputs from trainable neural networks

approximations then yields the corresponding distance metric. Complete proofs for all rows of Table 1 are provided in the Supplementary Section A.4.

Fig. 2 illustrates this correspondence for three representative velocity classes. A velocity field that is constant in x can only translate the Gaussian source. At the population optimum, this translation shifts the source mean to the target-distribution mean while leaving its covariance unchanged, so the Jeffreys divergence reduces to the squared Euclidean distance between target means. An affine-in-x velocity field can both translate and linearly transform the Gaussian source. Its population optimum therefore produces the moment-matched Gaussian approximation to the target distribution, and the resulting Jeffreys divergence is the Gaussian Jeffreys divergence. Finally, in the population limit, an unconstrained velocity field recovers the target distribution itself, yielding the full Jeffreys divergence.

Beyond a unification of existing distance metrics, the flow-matching framework enables principled designs of new metrics by specifying the source distribution as a distributional "template" and choosing a velocity class that constrains how the template can be transformed. The resulting distance is the Jeffreys divergence between the two fitted target distributions. As a demonstration, we design a new distance metric using a velocity class restricted to similarity transformations, thereby preserving the geometry of the source distribution when approximating the target distributions (Supplementary Section C). This metric can be useful when the geometry of the target distributions is known a priori or when one wishes to emphasize a particular geometric property.

5 Applications

We evaluate the flow-matching framework across a number of applications. First, we use a simple synthetic categorical dataset to test whether different velocity constraints recover multiple representational distance metrics (Section 5.1). The remaining four settings involve continuous variables: Fisher information in simulations and neural recordings (Section 5.2), and time-resolved RDMs in simulations and neural recordings (Section 5.3). The later four settings highlight an important advantage of flow matching for complex distributions that depend on continuous variables.

5.1 Flow matching recovers multiple distance metrics on a synthetic categorical dataset

We first test whether practically flow matching with different velocity constraints enables the estimation of commonly used representational distance metrics, as predicted by our analytical results. We construct a five-condition Gaussian dataset in $\mathbb{R}^3$, and fit flow matching with the corresponding velocity constraint for each metric in Table 1. Accuracy, measured by the Pearson correlation between the unique off-diagonal entries of the estimated and ground-truth RDMs, improves with total sample size across all six metrics (Fig. 3A). The estimated RDMs closely resemble the corresponding

ground-truth RDMs defined by the six metrics (Fig. 3B), supporting our main theoretical claim that flow matching can be used as a unified estimator for these metrics. Notably, most of these metrics depend only on first- and second-order moments, for which robust and efficient estimators such as Ledoit–Wolf shrinkage are available (Ledoit & Wolf, 2004). Indeed, when comparing flow matching with these specialized estimators on this simple Gaussian dataset, we did not observe advantages in estimated accuracy for flow matching (Fig. S1). In later applications, we will show that for datasets involving complex distributions and continuous variables, flow matching exhibits various advantages.

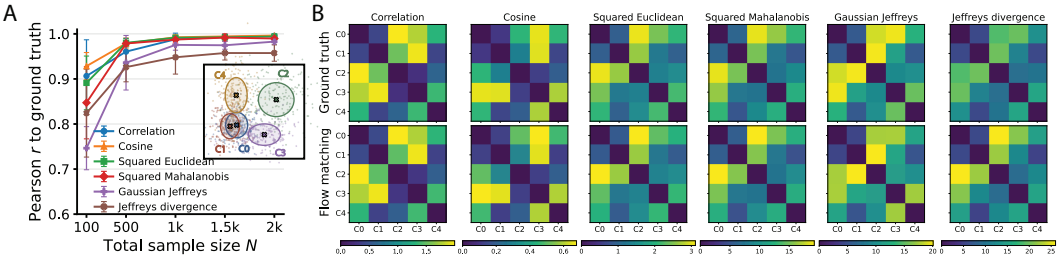

Figure 3: **Flow matching recovers distance metrics on a synthetic categorical dataset.** (**A**) Pearson correlation between the unique off-diagonal entries of the estimated and ground-truth RDMs as a function of total sample size N. Curves and error bars show means and sample standard deviations over five repetitions; the inset visualizes the five-condition dataset. (**B**) Ground-truth RDMs (top) and the corresponding flow-matching estimates using 2,000 total samples (bottom).

5.2 Flow matching estimates Fisher information in simulations and neural data

Fisher information is a local measure of representational distance. *Linear* Fisher information quantifies how precisely small stimulus changes can be estimated by an optimal, locally unbiased linear readout of neural responses (Brunel & Nadal, 1998; Abbott & Dayan, 1999; Averbeck et al., 2006; Moreno-Bote et al., 2014; Kanitscheider et al., 2015; Le & Wei, 2026), whereas *full* Fisher information measures the sensitivity of the entire response distribution through the log-likelihood score, including information accessible only through nonlinear readouts. Estimating Fisher information can be challenging, partly because it is difficult to model how response distributions depend on continuous conditions. In contrast, our flow-matching approach can, in principle, interpolate naturally across conditions by approximating the velocity field with a deep neural network, without requiring additional data binning or modeling assumptions. Below, we evaluate the performance of this flow-matching framework in estimating both quantities.

Simulations. We first estimate linear Fisher information using a toy dataset whose conditional response distribution is Gaussian given a continuous variable θ. We fit a condition-dependent affine (linear-in-state) velocity field, $v_\theta(x,t) = \dot{\beta}_t b_\theta + \bar{A}_{\theta,t}(x - \beta_t b_\theta)$ (Table 1). As baselines, we consider several existing methods. Local linear estimator (OLE) estimates linear Fisher information by binning the data and performing linear classification on adjacent bins (Kanitscheider et al., 2015). Gaussian-process kernel regression (GKR; see Section D.1.3) uses Gaussian-process regression to estimate the distributional mean as a function of θ and then uses kernel regression to estimate covariances (Ye & Wessel, 2025). The Wishart process jointly models the mean and covariance with Gaussian processes (Nejatbakhsh et al., 2023) (Section D.1.4). On this Gaussian dataset, affine flow yields a smooth estimate that closely follows the ground-truth linear Fisher-information when varying the sample-size and number of neurons(Fig. 4B, C). GKR and the Wishart process perform fairly well on low-dimensional datasets, but their errors increase substantially with response dimension (Fig. 4C; also see Pearson correlations in Supplementary Figure S2A).

Furthermore, we evaluate full Fisher information on another Gaussian mixture dataset (Fig. 4D). GKR and the Wishart process fail in estimation because they assume Gaussian distributions, whereas unconstrained flow successfully targets the full Fisher information (Fig. 4E and F).

Neural data. Next, we apply flow matching to neural recording data. We analyze six mouse visual-cortex recording sessions from Stringer et al. (2021), each containing responses from 10,000

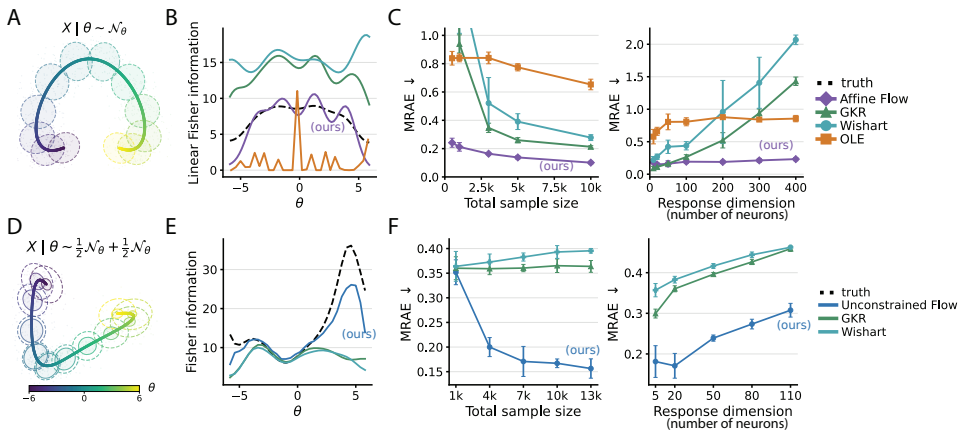

Figure 4: **Flow matching performs best in estimating Fisher information on toy datasets.** (**A**) Illustration of the Gaussian conditional-response dataset. (**B**) Representative linear Fisher-information estimates ($d = 300, N = 1000$). (**C**) Mean relative absolute error (MRAE) versus total sample size at $d = 300$ (left) and versus response dimension at $N = 1{,}000$ (right). (**D**) Equal-weight Gaussian scale-mixture dataset. (**E**) Representative full Fisher-information estimates. (**F**) MRAE versus total sample size at $d = 20$ (left) and versus response dimension at $N = 7{,}000$ (right). Points and error bars in (**C**) and (**F**) show means and standard deviations over five repetitions.

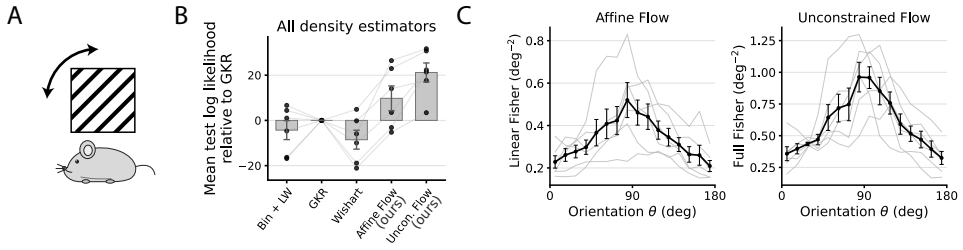

Figure 5: **Flow matching estimates Fisher information in mouse visual-cortex recordings.** (**A**) Experimental paradigm: visual-cortex responses are recorded while mice view static gratings. (**B**) Session-wise mean held-out log likelihood relative to GKR. Gray lines connect estimates from the same session. Bars and error bars show the across-session mean and standard error of the mean (SEM). (**C**) Linear Fisher information from affine flow and full Fisher information from unconstrained flow versus grating orientation. Gray lines show six individual sessions, each from a different mouse; black curves and error bars show the across-session mean and SEM.

to 25,000 neurons across about 4,000 static-grating trials (Fig. 5A). Direct estimation of Fisher information in this high-dimensional response space is statistically challenging and computationally demanding. Following common dimensionality-reduction practice (Cunningham & Yu, 2014), we performed PCA separately for each session using only the training trials and apply the fixed projection to the training, validation, and test responses (200 principal components, about 70% of the variance). All models are fitted and evaluated in this dimensionality-reduced PCA space.

For a given orientation range, visualizations of the PC1–PC2 subspace show that, while the Gaussian estimators (Bin + LW, GKR, the Wishart process, and affine flow) generally fit the held-out data well, some discrepancies remain (Supplementary Fig. S3). By contrast, unconstrained flow captures a non-elliptical density contour and visually fits the data better. Quantitatively, across the six sessions, affine flow has the highest mean held-out log likelihood among the Gaussian-family estimators. Furthermore, unconstrained flow, which does not impose Gaussianity, has the highest held-out log-likelihood overall in every session (Fig. 5B). Finally, the affine-flow linear Fisher estimate and the unconstrained-flow full Fisher estimate both peak near 90° (Fig. 5C; see similar Fisher curves estimated by the other methods in Supplementary Fig. S4), suggesting greater discriminability of horizontal visual patterns. These results are different from the oblique effects reported in humans (Appelle, 1972; Furmanski &

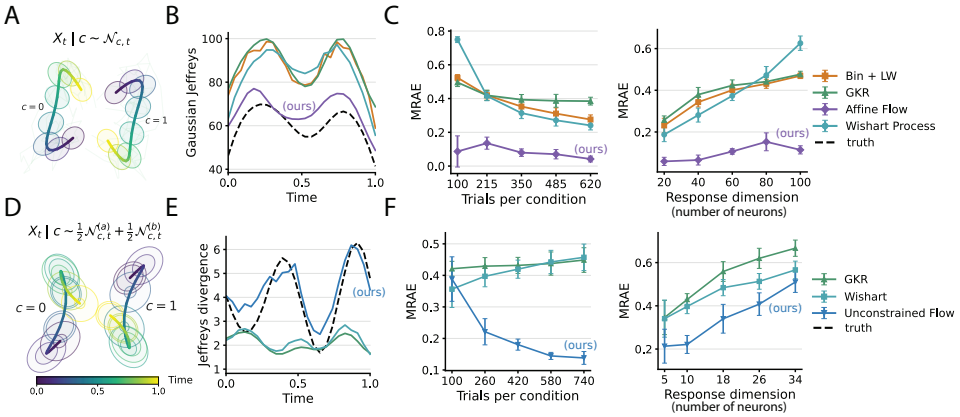

Figure 6: **Flow matching best estimates time-resolved distances on OU-trial toy datasets.** (**A**) Gaussian OU dataset with two conditions. (**B**) Representative time-resolved Gaussian Jeffreys-divergence estimates at $d = 60$ and 260 trials per condition. (**C**) MRAE versus trials per condition at $d = 60$ (left) and versus response dimension at 260 trials per condition (right). (**D**) Gaussian-mixture OU dataset. (**E**) Representative full Jeffreys-divergence estimates at $d = 10$ and 260 trials per condition. (**F**) MRAE versus trials per condition at $d = 10$ (left) and versus response dimension at 260 trials per condition (right). Points and error bars in (**C**) and (**F**) show means and standard deviations over five repetitions.

Engel, 2000) and macaques (Li et al., 2003), which show higher discriminability for both horizontal and vertical orientations, yet are consistent with a recent study in mice (Dipoppa et al., 2024).

5.3 Estimating time-resolved distances using flow matching

Neural population representations are not static, rather they evolve dynamically over time. Understanding such dynamics is important for understanding the computations performed by the brain. Time-resolved RDMs address this question by tracking how representational distances change over time. Conventional estimation typically involves dividing the data into discrete time bins. Our flow-matching framework does not require binning, as the velocity neural network learns to interpolate across time. In this section, we evaluate whether our method can estimate time-resolved RDMs.

Simulations. We construct two-condition toy datasets with temporally correlated Ornstein–Uhlenbeck (OU) trials (Uhlenbeck & Ornstein, 1930) and either Gaussian or two-scale Gaussian-mixture marginals. We measure pairwise distances using either Gaussian Jeffreys divergence or Jeffreys divergence. Gaussian Jeffreys divergence generalizes Mahalanobis distance by allowing the two conditions to have different covariance matrices, whereas Jeffreys divergence makes no Gaussian assumption and can therefore capture differences between their full distributions. We find that, on the Gaussian dataset, affine flow outperforms conventional methods in estimating Gaussian Jeffreys divergence throughout both the trial-count and response-dimension sweeps (Fig. 6A–C) On the mixture dataset, only the unconstrained flow-matching estimator captures the non-Gaussian structure and outperforms the other methods in estimating Jeffreys divergence (Fig. 6D–F).

Neural data. We next apply time-resolved distance estimation to extracellular spiking activity from the Allen Visual Coding Neuropixels dataset (Siegle et al., 2021). Each drifting-grating presentation is analyzed over four seconds: one second before stimulus onset, two seconds of stimulus presentation, and one second after stimulus offset (Fig. 7A). We treat each combination of motion direction and temporal frequency as a condition and examine how the pairwise representational distances evolve over time. We analyze 32 recording sessions, one for per mouse. Spikes are counted in nonoverlapping 100-ms bins, transformed elementwise using a variance-stablizing transformation as $k \mapsto \sqrt{k}$, and projected onto the top 70 principal components, which explains about 70% variance.

Quantitatively, affine flow leads to the highest mean held-out log-likelihood among the Gaussian estimators, whereas unconstrained flow has the highest mean held-out log-likelihood overall (Fig. 7B;

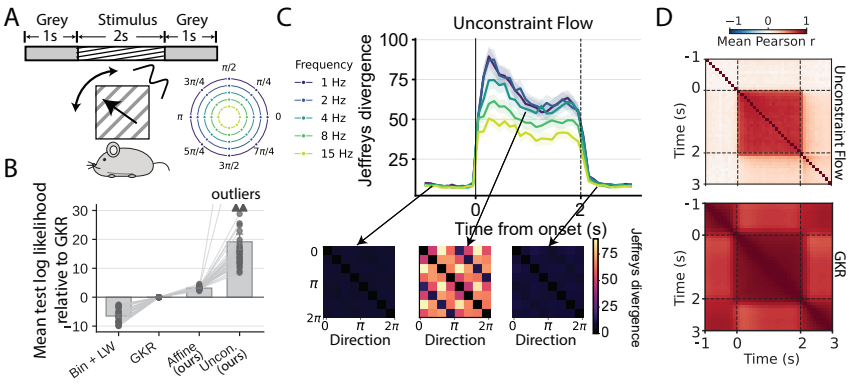

Figure 7: **Flow matching estimates time-resolved distances from Allen Neuropixels recordings.** (**A**) Four-second trial structure. (**B**) Mean test log-likelihood relative to the GKR method. Dots denote recording sessions, gray lines connect estimates from the same session, and bars show the mean across 32 sessions; triangles indicate two outliers above the displayed range. (**C**) Unconstrained-flow Jeffreys divergence averaged over the 28 unique pairs among the eight directions separately at each temporal frequency within each session (top), and 4-Hz direction RDMs averaged across sessions at the bin centers nearest -0.5, 1.0, and 2.5 seconds (bottom). Colors indicate temporal frequency. Curves and shaded bands show the across-session mean and SEM, respectively. (**D**) Time-to-time RDM correlations. Colors show mean Pearson correlations between RDMs at pairs of time points.

see example data visualizations in Supplementary Figs. S6 and S9). Similar results hold when using 10 rather than 70 principal components (Supplementary Fig. S10). Across all five temporal frequencies, both flow estimates of Jeffreys divergence rise sharply after stimulus onset and fall after stimulus offset. Higher temporal frequencies show smaller overall distances, suggesting poorer discrimination between grating directions and less directional information encoded at faster temporal frequencies (Fig. 7C; also see Supplementary Figs. S7 and S8). Individual RDMs exhibit semi-periodic structures, implying that therepresentational distances between drifting gratings of the opposite directions are closer compared to gratings with orthogonal directions (Fig. 7C bottom; also see Supplementary Fig. S11 for a similar pattern from analyzing calcium imaging data from de Vries et al. (2020)).

Using the RDMs at different time points, we can also assess neural representational stability over time by computing time-to-time RDM correlations (Fig. 7D; see Supplementary Fig. S13 for the other methods). The mean correlation is higher during stimulus presentation (0–2 s; mean = 0.73) than before (-1–0 s; mean = 0.055) or after (2–3 s; mean = 0.25). This suggests that the structure of neural representational geometry are relatively stable during stimulus presentation. Together with observations in Fig. 7C, these results suggest that, while the neural manifolds expand and shrink over time, the shape of the manifolds is generally preserved.

We also find that methods such as GKR tend to oversmooth the distribution, perhaps because they use Gaussian processes and assume Gaussian response distributions, producing spuriously high correlations of RDMs even before stimulus presentation (Fig. 7D).

6 Discussion and Conclusion

We have introduced a flow-matching framework that unifies neural representational dissimilarity metrics as Jeffreys divergences under different velocity-field constraints. It provides a common estimator with empirical advantages for high-dimensional data and complex distributions conditioned on continuous variables, and enables principled design of new metrics through source distributions and velocity constraints. Our results has several limitations. Because the theoretical correspondences hold at global population optima, which neural-network training may not reach, the fitted estimates can therefore deviate from their intended metric interpretations. Also, our Validation has focused on visual-cortex recordings. Extending it to other brain systems and measurement techniques (such as fMRI and EEG), as well as deep neural network models, remains an important direction for future investigations.

AI USE STATEMENT

AI-assisted tools were used for language editing and manuscript organization under author direction. The authors reviewed the scientific claims, derivations, citations, and conclusions and take responsibility for the final content.

ETHICS STATEMENT

This work analyzes synthetic data and previously collected animal recordings. No new animal or human-subject experiments were conducted for this study.

REPRODUCIBILITY STATEMENT

The Supplementary Material provides complete proofs, estimator specifications, dataset construction and preprocessing details, hyperparameter searches, and additional results. The real-data experiments use the visual-cortex recordings described by Stringer et al. (2021), two-photon calcium-imaging sessions from the Allen Brain Observatory Visual Coding dataset, and Allen Visual Coding Neuropixels recordings (Siegle et al., 2021). An anonymized code package will be provided with the submission.

REFERENCES

L. F. Abbott and Peter Dayan. The effect of correlated variability on the accuracy of a population code. *Neural Computation*, 11(1):91–101, 1999. doi: 10.1162/089976699300016827.

Michael S. Albergo and Eric Vanden-Eijnden. Building normalizing flows with stochastic interpolants. In *International Conference on Learning Representations*, 2023.

Shun-ichi Amari. *Information Geometry and Its Applications*. Springer Japan, 2016. doi: 10.1007/978-4-431-55978-8.

Stuart Appelle. Perception and discrimination as a function of stimulus orientation: the" oblique effect" in man and animals. *Psychological bulletin*, 78(4):266, 1972.

Bruno B. Averbeck, Peter E. Latham, and Alexandre Pouget. Neural correlations, population coding and computation. *Nature Reviews Neuroscience*, 7(5):358–366, 2006. doi: 10.1038/nrn1888.

Wessel Bruinsma, Eric Perim, William Tebbutt, Scott Hosking, Arno Solin, and Richard Turner. Scalable exact inference in multi-output gaussian processes. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1190–1201, 2020.

Nicolas Brunel and Jean-Pierre Nadal. Mutual information, fisher information, and population coding. *Neural Computation*, 10(7):1731–1757, 1998. doi: 10.1162/089976698300017115.

Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David K. Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems*, volume 31, 2018.

Radoslaw Martin Cichy, Dimitrios Pantazis, and Aude Oliva. Resolving human object recognition in space and time. *Nature Neuroscience*, 17:455–462, 2014. doi: 10.1038/nn.3635.

John P. Cunningham and Byron M. Yu. Dimensionality reduction for large-scale neural recordings. *Nature Neuroscience*, 17(11):1500–1509, 2014. doi: 10.1038/nn.3776.

Aram Davtyan, Sepehr Sameni, and Paolo Favaro. Efficient video prediction via sparsely conditioned flow matching. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 23263–23274, 2023.

Saskia E. J. de Vries et al. A large-scale standardized physiological survey reveals functional organization of the mouse visual cortex. *Nature Neuroscience*, 23:138–151, 2020. doi: 10.1038/s41593-019-0550-9.

Jörn Diedrichsen and Nikolaus Kriegeskorte. Representational models: A common framework for understanding encoding, pattern-component, and representational-similarity analysis. *PLOS Computational Biology*, 13(4):e1005508, 2017. doi: 10.1371/journal.pcbi.1005508.

Mario Dipoppa, Ramon Nogueira, Stéphane Bugeon, Yoni Friedman, Charu B. Reddy, Kenneth D. Harris, Dario L. Ringach, Kenneth D. Miller, Matteo Carandini, and Stefano Fusi. Adaptation shapes the representational geometry in mouse V1 to efficiently encode the environment. *bioRxiv*, 2024. doi: 10.1101/2024.12.11.628035.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas M"uller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *Proceedings of the 41st International Conference on Machine Learning*.

Ronald A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London. Series A*, 222:309–368, 1922. doi: 10.1098/rsta.1922.0009.

Christopher S. Furmanski and Stephen A. Engel. An oblique effect in human primary visual cortex. *Nature Neuroscience*, 3(6):535–536, 2000. doi: 10.1038/75702.

Will Grathwohl, Ricky T. Q. Chen, Jesse Bettencourt, Ilya Sutskever, and David K. Duvenaud. FFJORD: Free-form continuous dynamics for scalable reversible generative models. In *International Conference on Learning Representations*, 2019.

Brian C. Hall. *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*. Springer, 2 edition, 2015. doi: 10.1007/978-3-319-13467-3.

Vincent Hu, Di Wu, Yuki Asano, Pascal Mettes, Basura Fernando, Bj"orn Ommer, and Cees Snoek. Flow matching for conditional text generation in a few sampling steps. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 380–392. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.eacl-short.33.

Ingmar Kanitscheider, Ruben Coen-Cagli, Adam Kohn, and Alexandre Pouget. Measuring fisher information accurately in correlated neural populations. *PLOS Computational Biology*, 11(6): e1004218, 2015. doi: 10.1371/journal.pcbi.1004218.

Adam Kohn, Ruben Coen-Cagli, Ingmar Kanitscheider, and Alexandre Pouget. Correlations and neuronal population information. *Annual Review of Neuroscience*, 39:237–256, 2016. doi: 10.1146/annurev-neuro-070815-013851.

Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited. In *Proceedings of the 36th International Conference on Machine Learning*, pp. 3519–3529, 2019.

Nikolaus Kriegeskorte. Relating population-code representations between man, monkey, and computational models. *Frontiers in Neuroscience*, 3(3):363–373, 2009. doi: 10.3389/neuro.01.035.2009.

Nikolaus Kriegeskorte and Rogier A. Kievit. Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8):401–412, 2013. doi: 10.1016/j.tics.2013.06.007.

Nikolaus Kriegeskorte and Xue-Xin Wei. Neural tuning and representational geometry. *Nature Reviews Neuroscience*, 22(11):703–718, 2021. doi: 10.1038/s41583-021-00502-3.

Nikolaus Kriegeskorte, Marieke Mur, and Peter A. Bandettini. Representational similarity analysis: connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4, 2008a. doi: 10.3389/neuro.06.004.2008.

Nikolaus Kriegeskorte, Marieke Mur, Douglas A. Ruff, Roozbeh Kiani, Jerzy Bodurka, Hossein Esteky, Keiji Tanaka, and Peter A. Bandettini. Matching categorical object representations in inferior temporal cortex of man and monkey. *Neuron*, 60(6):1126–1141, 2008b. doi: 10.1016/j.neuron.2008.10.043.

Solomon Kullback and Richard A. Leibler. On information and sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86, 1951. doi: 10.1214/aoms/1177729694.

Dylan Le and Xue-Xin Wei. Split-trial analysis reveals the information capacity of neural population codes. *eLife*, 2026. doi: 10.7554/elife.111658.1.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4.

Baowang Li, Matthew R Peterson, and Ralph D Freeman. Oblique effect: a neural basis in the visual cortex. *Journal of neurophysiology*, 90(1):204–217, 2003.

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*, 2023.

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

Rubén Moreno-Bote, Jeffrey Beck, Ingmar Kanitscheider, Xaq Pitkow, Peter Latham, and Alexandre Pouget. Information-limiting correlations. *Nature Neuroscience*, 17(10):1410–1417, 2014. doi: 10.1038/nn.3807.

Amin Nejatbakhsh, Isabel Garon, and Alex Williams. Estimating noise correlations across continuous conditions with wishart processes. In *Advances in Neural Information Processing Systems*, volume 36, pp. 54032–54045, 2023.

Hamed Nili, Cai Wingfield, Alexander Walther, Li Su, William Marslen-Wilson, and Nikolaus Kriegeskorte. A toolbox for representational similarity analysis. *PLOS Computational Biology*, 10 (4):e1003553, 2014. doi: 10.1371/journal.pcbi.1003553.

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

Alexandre Pouget, Peter Dayan, and Richard Zemel. Information processing with population codes. *Nature Reviews Neuroscience*, 1(2):125–132, 2000. doi: 10.1038/35039062.

C. R. Rao. Information and accuracy attainable in the estimation of statistical parameters. *Bulletin of the Calcutta Mathematical Society*, 37(3):81–91, 1945.

Carl Edward Rasmussen and Christopher K. I. Williams. *Gaussian Processes for Machine Learning*. MIT Press, 2006. doi: 10.7551/mitpress/3206.001.0001.

Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning*, pp. 1530–1538, 2015.

Benjamin Rhodes, Kai Xu, and Michael U. Gutmann. Telescoping density-ratio estimation. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Peggy Seriès, Alan A. Stocker, and Eero P. Simoncelli. Is the homunculus "aware" of sensory adaptation? *Neural Computation*, 21(12):3271–3304, 2009. doi: 10.1162/neco.2009.09-08-869.

H. S. Seung and Haim Sompolinsky. Simple models for reading neuronal population codes. *Proceedings of the National Academy of Sciences*, 90(22):10749–10753, 1993. doi: 10.1073/pnas.90.22.10749.

Joshua H. Siegle et al. Survey of spiking in the mouse visual system reveals functional hierarchy. *Nature*, 592:86–92, 2021. doi: 10.1038/s41586-020-03171-x.

Carsen Stringer, Michalis Michaelos, Dmitri Tsyboulski, Sarah E. Lindo, and Marius Pachitariu. High-precision coding in visual cortex. *Cell*, 184(10):2767–2778.e15, 2021. doi: 10.1016/j.cell.2021.03.042.

George E. Uhlenbeck and Leonard S. Ornstein. On the theory of the brownian motion. *Physical Review*, 36(5):823–841, 1930. doi: 10.1103/PhysRev.36.823.

Alexander Walther, Hamed Nili, Naveed Ejaz, Arjen Alink, Nikolaus Kriegeskorte, and Jörn Diedrichsen. Reliability of dissimilarity measures for multi-voxel pattern analysis. *NeuroImage*, 137: 188–200, 2016. doi: 10.1016/j.neuroimage.2015.12.012.

Zeyuan Ye and Ralf Wessel. Speed modulations in grid cell information geometry. *Nature Communications*, 16:7723, 2025. doi: 10.1038/s41467-025-62856-x.

Supplementary Material

Supplementary Material Contents

A Mathematical Proofs of the Metric Correspondences under flow matching **15**
A.1 Assumptions and notation . . . 15
A.2 Affine-flow solution . . . 15
A.3 Deterministic Jeffreys evaluation for affine Gaussian flows . . . 15
A.4 Proofs of velocity-to-distance metric correspondences . . . 16

B Likelihood Fine-Tuning and Metric Correspondences **23**
B.1 Optional negative-log-likelihood fine-tuning . . . 23
B.2 Likelihood-based proofs . . . 24

C A new distance metric that preserves the geometry of a template **26**
C.1 Demonstration of a new distance metric that preserves the geometry of a template . 27

D Experimental Details **27**
D.1 Estimators . . . 27
D.1.1 Flow matching training . . . 27
D.1.2 Velocity network architecture . . . 28
D.1.3 Gaussian-process with kernel regression . . . 29
D.1.4 Wishart-process estimator . . . 29
D.1.5 Cross-fitted optimal linear estimator (OLE) . . . 31
D.1.6 TRE density-ratio estimation for categorical Jeffreys divergence . . . 31
D.2 Simulation datasets . . . 32
D.2.1 Categorical RDM dataset . . . 32
D.2.2 Fisher datasets . . . 32
D.2.3 Time-resolved Jeffreys datasets . . . 33
D.3 Real neural datasets . . . 34
D.3.1 Mouse V1 static-grating dataset . . . 34
D.3.2 Allen Brain Observatory two-photon calcium-imaging dataset with drifting gratings . . . 35
D.3.3 Allen Visual Coding Neuropixels dataset . . . 35
D.4 Estimator and dataset for the geometry-template Jeffreys distance . . . 36

E Supplementary Figures **37**

A Mathematical Proofs of the Metric Correspondences under Flow Matching

In this section, we show that flow-matching training under different velocity constraints yields Jeffreys divergences that correspond to different existing distance metrics. Optional log-likelihood fine-tuning recovers the same distance metrics as well, as shown in Supplementary Section B.

A.1 Assumptions and notation

We use the term representational *distance* in the sense common in neural data analysis: it is symmetric and nonnegative but need not obey the triangle inequality.

We prove the flow-matching population statements in Table 1. Unless stated otherwise, the source is $\rho_0 = \mathcal{N}(0, I)$, $X_0 \sim \rho_0$, $X_1^{(c)} \sim p_c$, and X_0 is independent of $X_1^{(c)}$. Each modeled condition has $\pi_{\mathcal{C}}(c) > 0$, and each p_c has finite second moments. Covariance matrices are invertible. The schedules are continuously differentiable and satisfy $\alpha_0 = 1$, $\alpha_1 = 0$, $\beta_0 = 0$, and $\beta_1 = 1$. For condition c, write

$$X_t^{(c)} = \alpha_t X_0 + \beta_t X_1^{(c)}, \qquad U_t^{(c)} = \dot{\alpha}_t X_0 + \dot{\beta}_t X_1^{(c)}. \tag{9}$$

For a velocity class $\mathcal{V}$, let $v^{\mathcal{V}} = \arg\min_{v \in \mathcal{V}} \mathcal{L}_{\mathrm{FM}}(v)$, with $\mathcal{L}_{\mathrm{FM}}$ defined in Eq. (5). Inserting $v^{\mathcal{V}}$ into the ODE produces $q_c^{\mathcal{V}}$ for each generic condition c. Pairwise results compare $q_a^{\mathcal{V}}$ and $q_b^{\mathcal{V}}$. The unknown response distributions p_c need not be Gaussian.

Before proving the correspondences, we provide two formulas used repeatedly below: the affine-flow endpoint distribution obtained by setting $t = 1$ in Eq. (12) (Subsection A.2) and the closed-form Jeffreys divergence between Gaussian endpoints in Eq. (16) (Subsection A.3).

A.2 Affine-flow solution

Lemma A.1 (Affine flows preserve Gaussianity)**.** *Let $A_t \in \mathbb{R}^{d \times d}$ and $r_t \in \mathbb{R}^d$ be continuous on $[0, 1]$, and consider the affine ODE*

$$\dot{X}_t = A_t X_t + r_t, \qquad X_0 \sim \mathcal{N}(m_0, S_0). \tag{10}$$

Let Φ_t solve $\dot{\Phi}_t = A_t \Phi_t$ with $\Phi_0 = I$, and let η_t solve $\dot{\eta}_t = A_t \eta_t + r_t$ with $\eta_0 = 0$. Then

$$X_t = \Phi_t X_0 + \eta_t \tag{11}$$

and therefore

$$X_t \sim \mathcal{N}\big(\Phi_t m_0 + \eta_t, \Phi_t S_0 \Phi_t^\top\big) \tag{12}$$

for every $t \in [0, 1]$.

Proof. Differentiating the right-hand side of Eq. (11) shows that it satisfies the affine ODE and the initial condition X_0. Uniqueness of the linear ODE therefore gives Eq. (11). This is an affine transformation of the Gaussian random vector X_0, so Eq. (12) follows. □

A.3 Deterministic Jeffreys evaluation for affine Gaussian flows

Suppose the fitted velocity for condition c is affine in the state,

$$v_c^{\mathcal{V}}(x, t) = A_{c,t} x + r_{c,t}, \tag{13}$$

and the source is $\rho_0 = \mathcal{N}(m_0, S_0)$. By Lemma A.1, the distribution remains Gaussian along the flow. Its mean $m_{c,t} = \mathbb{E}[X_t]$ and covariance $S_{c,t} = \mathrm{Cov}(X_t)$ satisfy

$$\begin{aligned} \dot{m}_{c,t} &= A_{c,t} m_{c,t} + r_{c,t}, & m_{c,0} &= m_0, \\ \dot{S}_{c,t} &= A_{c,t} S_{c,t} + S_{c,t} A_{c,t}^\top, & S_{c,0} &= S_0. \end{aligned} \tag{14}$$

The first equation follows by taking the expectation of $\dot{X}_t = A_{c,t} X_t + r_{c,t}$. For the second, write $Y_t = X_t - m_{c,t}$, so that $\dot{Y}_t = A_{c,t} Y_t$, and differentiate $S_{c,t} = \mathbb{E}[Y_t Y_t^\top]$. Integrating Eq. (14) from $t = 0$ to $t = 1$ therefore gives the endpoint distribution directly:

$$q_c^{\mathcal{V}} = \mathcal{N}(m_c, S_c), \qquad m_c = m_{c,1}, \quad S_c = S_{c,1}. \tag{15}$$

For two conditions a and b, let $\delta_{ab} = m_a - m_b$. Their Jeffreys divergence is then available in closed form:

$$D_{\mathcal{V}}(a,b) = \frac{1}{2}\left[\mathrm{tr}(S_b^{-1}S_a) + \mathrm{tr}(S_a^{-1}S_b) - 2d + \delta_{ab}^{\top}(S_a^{-1} + S_b^{-1})\delta_{ab}\right]. \tag{16}$$

In practice, we solve Eq. (14) once per condition, symmetrize the numerical endpoint covariance as $(S_c + S_c^{\top})/2$, and evaluate Eq. (16) using Cholesky-based linear solves rather than explicit matrix inverses. All entries of an RDM can then be formed from the stored endpoint moments. Unlike Eq. (8), this procedure of Jeffreys divergence computation requires no generated samples or sample-wise log-density evaluations and has no Monte Carlo sampling error.

A.4 Proofs of velocity-to-distance metric correspondences

Proposition A.2 (Translation-constrained flow matching recovers squared Euclidean distance)**.** *Let $\rho_0 = \mathcal{N}(0, I)$, let p_c have finite mean μ_c, assume $\pi_{\mathcal{C}}(c) > 0$, and let β_t be continuously differentiable with $\beta_0 = 0$, $\beta_1 = 1$, and $K_\beta = \int_0^1 \dot\beta_t^2\, dt > 0$. For the translation-only velocity class*

$$\mathcal{V}_{\mathrm{Euc}} = \{v_c(x,t) = \dot\beta_t b_c : b_c \in \mathbb{R}^d\}. \tag{17}$$

Write $\mathbf{b} = (b_c)_{c\in\mathcal{C}}$ and let $v_{\mathbf{b}}$ denote the corresponding condition-dependent velocity field. The global minimizer of the population flow-matching loss in Eq. (5),

$$\mathbf{b}^\star = \arg\min_{\mathbf{b}} \mathcal{L}_{\mathrm{FM}}(v_{\mathbf{b}}), \tag{18}$$

has condition-specific components $b_c^\star = \mu_c$ and produces $q_c^{\mathcal{V}_{\mathrm{Euc}}} = \mathcal{N}(\mu_c, I)$. Consequently,

$$D_{\mathcal{V}_{\mathrm{Euc}}}(a,b) = \|\mu_a - \mu_b\|_2^2. \tag{19}$$

Proof. Expanding the objective to the sampled condition C gives

$$\mathcal{L}_{\mathrm{FM}}(v_{\mathbf{b}}) = \mathbb{E}_{C\sim\pi_{\mathcal{C}}}\left[\mathcal{L}_{\mathrm{FM},C}(b_C)\right] = \sum_{c\in\mathcal{C}} \pi_{\mathcal{C}}(c)\, \mathcal{L}_{\mathrm{FM},c}(b_c), \tag{20}$$

where $\mathcal{L}_{\mathrm{FM},c}$ is the flow-matching loss conditional on $C = c$. Each b_c appears in only one summand of Eq. (20). Because $\pi_{\mathcal{C}}(c) > 0$, the global minimization therefore separates as

$$b_c^\star = \arg\min_{b_c\in\mathbb{R}^d} \mathcal{L}_{\mathrm{FM},c}(b_c) \qquad \text{for every } c \in \mathcal{C}. \tag{21}$$

We can now solve each condition-specific problem. Substituting $v_c(x,t) = \dot\beta_t b_c$ and $U_t^{(c)} = \dot\alpha_t X_0 + \dot\beta_t X_1^{(c)}$ gives

$$\mathcal{L}_{\mathrm{FM},c}(b_c) = \int_0^1 \mathbb{E}\left[\left\|\dot\beta_t b_c - \dot\alpha_t X_0 - \dot\beta_t X_1^{(c)}\right\|_2^2 \middle| c\right] dt \tag{22}$$

$$= \int_0^1 \Big[\dot\beta_t^2\|b_c\|_2^2 - 2\dot\alpha_t\dot\beta_t b_c^{\top}\mathbb{E}[X_0] - 2\dot\beta_t^2 b_c^{\top}\mathbb{E}[X_1^{(c)} \mid c] + \mathbb{E}\left[\left\|\dot\alpha_t X_0 + \dot\beta_t X_1^{(c)}\right\|_2^2 \middle| c\right]\Big] dt. \tag{23}$$

The final expectation in Eq. (23) does not depend on b_c. Since $X_0 \sim \mathcal{N}(0, I)$ and $X_1^{(c)} \sim p_c$, we have $\mathbb{E}[X_0] = 0$ and $\mathbb{E}[X_1^{(c)} \mid c] = \mu_c$. Therefore

$$\mathcal{L}_{\mathrm{FM},c}(b_c) = K_\beta\|b_c\|_2^2 - 2K_\beta b_c^{\top}\mu_c + \mathrm{const} \tag{24}$$

$$= K_\beta\|b_c - \mu_c\|_2^2 + \mathrm{const}, \tag{25}$$

where $K_\beta = \int_0^1 \dot\beta_t^2\, dt > 0$. Hence $b_c^\star = \mu_c$. The time-one flow is the translation $x_1 = x_0 + \mu_c$, so Lemma A.1 gives $q_c^{\mathcal{V}_{\mathrm{Euc}}} = \mathcal{N}(\mu_c, I)$. Applying this result to conditions a and b, their Jeffreys divergence is

$$D_{\mathcal{V}_{\mathrm{Euc}}}(a,b) = \|\mu_a - \mu_b\|_2^2. \tag{26}$$

Note that the response distribution p_c need not be Gaussian: the Gaussian form of $q_c^{\mathcal{V}_{\mathrm{Euc}}}$ follows from translating the Gaussian source. □

Proposition A.3 (Fixed-norm translations give cosine distance). *Let $r > 0$, suppose $K_\beta = \int_0^1 \dot{\beta}_t^2 \, dt > 0$, and suppose $\mu_c \neq 0$ for every condition. For the fixed-norm translation class*

$$\mathcal{V}_{\cos} = \{v_c(x,t) = \dot{\beta}_t b_c : \ \|b_c\| = r\}. \tag{27}$$

Then the population flow-matching optimum is unique within each condition and satisfies $b_c^\star = r\mu_c/\|\mu_c\|$. It produces

$$q_c^{\mathcal{V}_{\cos}} = \mathcal{N}\left(r\frac{\mu_c}{\|\mu_c\|}, I\right), \tag{28}$$

and the induced distance is

$$D_{\mathcal{V}_{\cos}}(a,b) = 2r^2\left(1 - \frac{\mu_a^\top \mu_b}{\|\mu_a\|\|\mu_b\|}\right), \tag{29}$$

which is $2r^2$ times cosine distance between the condition means.

Proof. As in Proposition A.2, the global objective separates over conditions. Restricting Eq. (25) to $\|b_c\| = r$ makes $K_\beta\|b_c\|^2$ constant, so minimizing the condition-specific loss is equivalent to

$$\max_{\|b_c\|=r} b_c^\top \mu_c. \tag{30}$$

By Cauchy–Schwarz, $b_c^\top \mu_c \leq r\|\mu_c\|$, with equality only when b_c points in the direction of μ_c. Hence

$$b_c^\star = r\frac{\mu_c}{\|\mu_c\|}. \tag{31}$$

The endpoint ODE integrates to $x_1 = x_0 + b_c^\star$. By Lemma A.1,

$$q_c^{\mathcal{V}_{\cos}} = \mathcal{N}\left(r\frac{\mu_c}{\|\mu_c\|}, I\right). \tag{32}$$

Writing $u_c = \mu_c/\|\mu_c\|$, Jeffreys divergence between the fitted distributions for a and b is

$$D_{\mathcal{V}_{\cos}}(a,b) = \|ru_a - ru_b\|^2 = 2r^2(1 - u_a^\top u_b) \tag{33}$$

$$= 2r^2\left(1 - \frac{\mu_a^\top \mu_b}{\|\mu_a\|\|\mu_b\|}\right). \tag{34}$$

Thus the constraint removes mean-magnitude information and retains only direction. If $\mu_c = 0$, every point on the radius-r sphere is optimal, so neither the fitted direction nor cosine distance is defined uniquely. □

Proposition A.4 (Centered fixed-norm translations give correlation distance). *Let $H = I - \mathbf{1}\mathbf{1}^\top/d$, let $r > 0$, suppose $K_\beta > 0$, and suppose $H\mu_c \neq 0$ for every condition. For the centered fixed-norm class*

$$\mathcal{V}_{\text{corr}} = \{v_c(x,t) = \dot{\beta}_t b_c : \ \mathbf{1}^\top b_c = 0, \ \|b_c\| = r\}. \tag{35}$$

Then the population flow-matching optimum is unique within each condition and satisfies $b_c^\star = rH\mu_c/\|H\mu_c\|$. It produces

$$q_c^{\mathcal{V}_{\text{corr}}} = \mathcal{N}\left(r\frac{H\mu_c}{\|H\mu_c\|}, I\right), \tag{36}$$

and the induced distance is

$$D_{\mathcal{V}_{\text{corr}}}(a,b) = 2r^2\left(1 - \frac{(H\mu_a)^\top (H\mu_b)}{\|H\mu_a\|\|H\mu_b\|}\right), \tag{37}$$

which is $2r^2$ times Pearson correlation distance between the condition means.

Proof. The matrix H is the orthogonal projector onto the centered subspace $\{z : \mathbf{1}^\top z = 0\}$: $H^\top = H$, $H^2 = H$, and $H\mathbf{1} = 0$. Proposition A.3 shows that, under a fixed-norm translation constraint, minimizing the flow-matching loss is equivalent to maximizing $b_c^\top \mu_c$. The additional

constraint in Eq. (35) restricts this maximization to the centered sphere. Every feasible b_c satisfies $b_c = Hb_c$, so

$$b_c^\top \mu_c = b_c^\top H \mu_c. \tag{38}$$

Therefore

$$b_c^\top \mu_c = b_c^\top H \mu_c \le r\|H\mu_c\|, \tag{39}$$

with equality only when b_c is aligned with $H\mu_c$. Thus

$$b_c^\star = r\frac{H\mu_c}{\|H\mu_c\|}. \tag{40}$$

Applying Lemma A.1 gives

$$q_c^{\mathcal{V}_{\text{corr}}} = \mathcal{N}\left(r\frac{H\mu_c}{\|H\mu_c\|}, I\right). \tag{41}$$

Writing $\widetilde{u}_c = H\mu_c/\|H\mu_c\|$, the Jeffreys divergence is

$$D_{\mathcal{V}_{\text{corr}}}(a,b) = \|r\widetilde{u}_a - r\widetilde{u}_b\|^2 \tag{42}$$

$$= 2r^2\left(1 - \frac{(H\mu_a)^\top (H\mu_b)}{\|H\mu_a\|\|H\mu_b\|}\right). \tag{43}$$

The inner product is Pearson correlation after feature-wise centering and normalization. If $H\mu_c = 0$, the constrained optimum is nonunique and correlation distance is undefined for that condition. □

Proposition A.5 (A shared centered affine flow gives Mahalanobis distance)**.** *Let* $\Sigma_c = \mathrm{Cov}_{p_c}(X)$ *and* $\Sigma_{\text{pool}} = \mathbb{E}_{C\sim\pi_C}[\Sigma_C]$. *Consider*

$$\mathcal{V}_{\text{Mah}} = \{v_c(x,t) = \dot\beta_t b_c + A_t(x - \beta_t b_c)\}, \tag{44}$$

where b_c *is condition specific and* A_t *is shared across conditions. Define*

$$G(A) = \int_0^1 (\dot\beta_t I - \beta_t A_t)^\top (\dot\beta_t I - \beta_t A_t)\, dt. \tag{45}$$

Assume the pooled path covariance $\bar{S}_t = \alpha_t^2 I + \beta_t^2 \Sigma_{\text{pool}}$ *is invertible for every* t, *and assume* $G(A^\star)$ *is positive definite at a population optimum. Then every population optimum has* $b_c^\star = \mu_c$ *and induces*

$$q_c^{\mathcal{V}_{\text{Mah}}} = \mathcal{N}(\mu_c, \Sigma_{\text{pool}}). \tag{46}$$

Consequently,

$$D_{\mathcal{V}_{\text{Mah}}}(a,b) = (\mu_a - \mu_b)^\top \Sigma_{\text{pool}}^{-1}(\mu_a - \mu_b), \tag{47}$$

which is the squared Mahalanobis distance based on pooled within-condition covariance.

Proof. Write

$$X_1^{(c)} = \mu_c + \varepsilon_c, \qquad \mathbb{E}[\varepsilon_c \mid c] = 0, \qquad \mathrm{Cov}(\varepsilon_c \mid c) = \Sigma_c. \tag{48}$$

where ε_c does not need to be Gaussian. Define the centered path variables

$$Y_t^{(c)} = \alpha_t X_0 + \beta_t \varepsilon_c, \qquad W_t^{(c)} = \dot\alpha_t X_0 + \dot\beta_t \varepsilon_c. \tag{49}$$

Then $X_t^{(c)} = \beta_t\mu_c + Y_t^{(c)}$ and $U_t^{(c)} = \dot\beta_t\mu_c + W_t^{(c)}$, with both $Y_t^{(c)}$ and $W_t^{(c)}$ centered. Substitution gives

$$v_c(X_t^{(c)}, t) - U_t^{(c)} = A_t Y_t^{(c)} - W_t^{(c)} + (\dot\beta_t I - \beta_t A_t)(b_c - \mu_c). \tag{50}$$

Expanding the squared velocity error using Eq. (50) produces the cross term

$$2\big(A_t Y_t^{(c)} - W_t^{(c)}\big)^\top (\dot\beta_t I - \beta_t A_t)(b_c - \mu_c). \tag{51}$$

Conditioned on c and t, the second factor is deterministic, whereas $\mathbb{E}[A_t Y_t^{(c)} - W_t^{(c)} \mid c, t] = 0$ because both $Y_t^{(c)}$ and $W_t^{(c)}$ are centered. The expected cross term therefore vanishes, and the population loss decomposes into

$$\mathcal{L}_{\text{FM}}(A, \mathbf{b}) = \mathbb{E}_{\substack{C\sim\pi_C \\ t\sim\mathrm{Unif}[0,1]}}\Big[\mathbb{E}\Big[\|A_t Y_t^{(C)} - W_t^{(C)}\|^2 \mid C, t\Big] + \|(\dot\beta_t I - \beta_t A_t)(b_C - \mu_C)\|^2\Big]. \tag{52}$$

For fixed A, the first term does not depend on b_c. The second term is nonnegative; choosing $b_c = \mu_c$ makes it exactly zero. Thus $b_c^\star = \mu_c$ minimizes the loss over b_c for every fixed A.

It remains to identify the shared covariance learned by A_t. Let

$$\Sigma_{\text{pool}} = \mathbb{E}_{C\sim\pi_C}[\Sigma_C], \qquad \bar{S}_t = \mathbb{E}_{C\sim\pi_C}[\text{Cov}(Y_t^{(C)} \mid C)] = \alpha_t^2 I + \beta_t^2 \Sigma_{\text{pool}}, \tag{53}$$

and

$$\bar{R}_t = \mathbb{E}_{C\sim\pi_C}[\text{Cov}(W_t^{(C)}, Y_t^{(C)} \mid C)] = \alpha_t\dot{\alpha}_t I + \beta_t\dot{\beta}_t\Sigma_{\text{pool}}. \tag{54}$$

For each fixed t, differentiating the centered least-squares objective with respect to A_t gives the normal equation

$$A_t^\star \bar{S}_t = \bar{R}_t. \tag{55}$$

Because $\bar{S}_t$ is invertible, $A_t^\star = \bar{R}_t\bar{S}_t^{-1}$ is unique.

Let $m_{c,t} = \mathbb{E}[x_t \mid c]$ denote the mean generated by the learned affine flow. Substituting $b_c^\star = \mu_c$ and $A_t^\star$ into Eq. (1) and taking the conditional expectation of both sides gives

$$\dot{m}_{c,t} = \dot{\beta}_t\mu_c + A_t^\star(m_{c,t} - \beta_t\mu_c), \qquad m_{c,0} = 0. \tag{56}$$

Here we used $\mathbb{E}[A_t^\star(x_t - \beta_t\mu_c) \mid c] = A_t^\star(m_{c,t} - \beta_t\mu_c)$ because $A_t^\star$ and μ_c are fixed for a given t and condition c. The function $m_{c,t} = \beta_t\mu_c$ solves this ODE, so uniqueness gives $m_{c,1} = \mu_c$. The generated covariance Q_t satisfies

$$\frac{dQ_t}{dt} = A_t^\star Q_t + Q_t A_t^{\star\top}, \qquad Q_0 = I. \tag{57}$$

But

$$\frac{d\bar{S}_t}{dt} = 2\alpha_t\dot{\alpha}_t I + 2\beta_t\dot{\beta}_t\Sigma_{\text{pool}} = \bar{R}_t + \bar{R}_t^\top = A_t^\star\bar{S}_t + \bar{S}_t A_t^{\star\top}. \tag{58}$$

Thus $\bar{S}_t$ solves the same covariance ODE with the same initial condition. By the uniqueness theorem for linear ODEs, $Q_t = \bar{S}_t$ and $Q_1 = \Sigma_{\text{pool}}$. Lemma A.1 shows that the endpoint distribution is Gaussian; together with the mean and covariance derived above, this gives

$$q_c^{\mathcal{V}_{\text{Mah}}} = \mathcal{N}(\mu_c, \Sigma_{\text{pool}}). \tag{59}$$

Applying the Gaussian KL formula to conditions a and b gives

$$D_{\mathcal{V}_{\text{Mah}}}(a, b) = (\mu_a - \mu_b)^\top\Sigma_{\text{pool}}^{-1}(\mu_a - \mu_b). \tag{60}$$

This conclusion depends only on the first two moments of each p_c; the true response distributions need not be Gaussian. □

Proposition A.6 (Condition-specific affine flows give Gaussian Jeffreys divergence)**.** *Let $\Sigma_c = \text{Cov}_{p_c}(X)$ be positive definite. For the condition-specific centered affine class*

$$\mathcal{V}_{\text{Aff}} = \{v_c(x,t) = \dot{\beta}_t b_c + A_{c,t}(x - \beta_t b_c)\}, \tag{61}$$

define

$$S_{c,t} = \alpha_t^2 I + \beta_t^2\Sigma_c, \qquad G_c(A_c) = \int_0^1 (\dot{\beta}_t I - \beta_t A_{c,t})^\top(\dot{\beta}_t I - \beta_t A_{c,t})\, dt. \tag{62}$$

Assume $S_{c,t}$ is invertible for every (c,t) and $G_c(A_c^\star)$ is positive definite at a population optimum. Then $b_c^\star = \mu_c$ and

$$q_c^{\mathcal{V}_{\text{Aff}}} = \mathcal{N}(\mu_c, \Sigma_c). \tag{63}$$

For $\delta_{ab} = \mu_a - \mu_b$, the induced distance is

$$D_{\mathcal{V}_{\text{Aff}}}(a, b) = \frac{1}{2}\Big[\,\text{tr}(\Sigma_b^{-1}\Sigma_a) + \text{tr}(\Sigma_a^{-1}\Sigma_b) - 2d + \delta_{ab}^\top(\Sigma_a^{-1} + \Sigma_b^{-1})\delta_{ab}\Big]. \tag{64}$$

Proof. Because both b_c and $A_{c,t}$ are condition specific, Eq. (20) separates the global optimization into independent condition-specific problems. Write $X_1^{(c)} = \mu_c + \varepsilon_c$ and define $Y_t^{(c)}$ and $W_t^{(c)}$ as in the Mahalanobis proof. Then

$$v_c(X_t^{(c)}, t) - U_t^{(c)} = A_{c,t}Y_t^{(c)} - W_t^{(c)} + (\dot{\beta}_t I - \beta_t A_{c,t})(b_c - \mu_c). \tag{65}$$

The condition-specific loss decomposes into

$$\mathcal{L}_{\mathrm{FM},c}(A_c, b_c) = \int_0^1 \mathbb{E}\Big[\|A_{c,t}Y_t^{(c)} - W_t^{(c)}\|^2 \mid c\Big]\, dt + (b_c - \mu_c)^\top G_c(A_c)(b_c - \mu_c). \tag{66}$$

For fixed A_c, the first term is independent of b_c, whereas the second term is nonnegative and vanishes at $b_c = \mu_c$. Hence $b_c = \mu_c$ minimizes the condition-specific loss for every fixed A_c. Positive definiteness of $G_c(A_c^\star)$ makes this minimizer unique at the population optimum, so $b_c^\star = \mu_c$.

For the centered problem, let

$$R_{c,t} = \mathrm{Cov}(W_t^{(c)}, Y_t^{(c)} \mid c) = \alpha_t\dot{\alpha}_t I + \beta_t\dot{\beta}_t\Sigma_c. \tag{67}$$

Differentiating the least-squares loss with respect to $A_{c,t}$ gives

$$A_{c,t}^\star S_{c,t} = R_{c,t}, \qquad A_{c,t}^\star = R_{c,t}S_{c,t}^{-1}. \tag{68}$$

As above, the generated mean is $m_{c,t} = \beta_t\mu_c$. Its covariance $Q_{c,t}$ solves

$$\dot{Q}_{c,t} = A_{c,t}^\star Q_{c,t} + Q_{c,t}A_{c,t}^{\star\top}, \qquad Q_{c,0} = I. \tag{69}$$

Meanwhile,

$$\dot{S}_{c,t} = R_{c,t} + R_{c,t}^\top = A_{c,t}^\star S_{c,t} + S_{c,t}A_{c,t}^{\star\top}, \qquad S_{c,0} = I. \tag{70}$$

Uniqueness of the covariance ODE gives $Q_{c,t} = S_{c,t}$, and hence $Q_{c,1} = \Sigma_c$. By Lemma A.1, the endpoint distribution is Gaussian, so the affine flow maps ρ_0 to

$$q_c^{\mathcal{V}_{\mathrm{Aff}}} = \mathcal{N}(\mu_c, \Sigma_c). \tag{71}$$

Therefore,

$$D_{\mathcal{V}_{\mathrm{Aff}}}(a, b) = \frac{1}{2}\big[\,\mathrm{tr}(\Sigma_b^{-1}\Sigma_a) + \mathrm{tr}(\Sigma_a^{-1}\Sigma_b) - 2d + \delta_{ab}^\top(\Sigma_a^{-1} + \Sigma_b^{-1})\delta_{ab}\big]. \tag{72}$$

When $\Sigma_a = \Sigma_b = \Sigma$, the trace terms cancel and the result reduces to $(\mu_a - \mu_b)^\top\Sigma^{-1}(\mu_a - \mu_b)$. □

Proposition A.7 (Stimulus-dependent affine flows give Gaussian Fisher information)**.** *Let θ be a scalar condition and suppose*

$$q_\theta^{\mathcal{V}_{\mathrm{Aff}}} = \mathcal{N}(\mu_\theta, \Sigma_\theta). \tag{73}$$

Assume μ_θ and Σ_θ are twice continuously differentiable in a neighborhood of θ and Σ_θ is positive definite there. Then, as $h \to 0$,

$$D_{\mathrm{J}}\Big(q_\theta^{\mathcal{V}_{\mathrm{Aff}}}, q_{\theta+h}^{\mathcal{V}_{\mathrm{Aff}}}\Big) = J_{\mathrm{Gauss}}(\theta)h^2 + O(h^3), \tag{74}$$

where

$$J_{\mathrm{Gauss}}(\theta) = \mu_\theta'^\top\Sigma_\theta^{-1}\mu_\theta' + \frac{1}{2}\,\mathrm{tr}\big(\Sigma_\theta^{-1}\Sigma_\theta'\Sigma_\theta^{-1}\Sigma_\theta'\big)\,. \tag{75}$$

Proof. Applying Proposition A.6 to θ and $\theta+h$, write $D_h = D_{\mathrm{J}}(q_\theta^{\mathcal{V}_{\mathrm{Aff}}}, q_{\theta+h}^{\mathcal{V}_{\mathrm{Aff}}})$ and $\Delta\mu_h = \mu_\theta - \mu_{\theta+h}$. Then

$$D_h = \underbrace{\frac{1}{2}\left[\mathrm{tr}(\Sigma_{\theta+h}^{-1}\Sigma_\theta) + \mathrm{tr}(\Sigma_\theta^{-1}\Sigma_{\theta+h}) - 2d\right]}_{\text{covariance contribution}} + \underbrace{\frac{1}{2}\Delta\mu_h^\top(\Sigma_\theta^{-1} + \Sigma_{\theta+h}^{-1})\Delta\mu_h}_{\text{mean-difference contribution}}\,. \tag{76}$$

Taylor expansion gives

$$\mu_{\theta+h} = \mu_\theta + \mu_\theta' h + O(h^2), \qquad \Sigma_{\theta+h} = \Sigma_\theta + \Sigma_\theta' h + O(h^2). \tag{77}$$

Because $\Sigma_{\theta+h}^{-1} = \Sigma_\theta^{-1} + O(h)$, the mean-difference contribution is

$$\mu_\theta'^\top\Sigma_\theta^{-1}\mu_\theta'\, h^2 + O(h^3). \tag{78}$$

For the covariance contribution, define

$$B = \Sigma_\theta^{-1/2}(\Sigma_{\theta+h} - \Sigma_\theta)\Sigma_\theta^{-1/2} = \Sigma_\theta^{-1/2}\Sigma_\theta'\Sigma_\theta^{-1/2}h + O(h^2). \tag{79}$$

From the definition of B, we have $\Sigma_{\theta+h} = \Sigma_\theta^{1/2}(I+B)\Sigma_\theta^{1/2}$. Hence $\Sigma_\theta^{-1}\Sigma_{\theta+h} = \Sigma_\theta^{-1/2}(I+B)\Sigma_\theta^{1/2}$ is similar to $I+B$, while $\Sigma_{\theta+h}^{-1}\Sigma_\theta = \Sigma_\theta^{-1/2}(I+B)^{-1}\Sigma_\theta^{1/2}$ is similar to $(I+B)^{-1}$. Similarity preserves trace, so the covariance trace can be evaluated using $I+B$ and $(I+B)^{-1}$. Since $(I+B)^{-1} = I - B + B^2 + O(\|B\|^3)$, the first-order terms cancel and

$$\frac{1}{2}\operatorname{tr}\left(\Sigma_{\theta+h}^{-1}\Sigma_\theta + \Sigma_\theta^{-1}\Sigma_{\theta+h} - 2I\right) = \frac{1}{2}\operatorname{tr}\left(\Sigma_\theta^{-1}\Sigma_\theta'\Sigma_\theta^{-1}\Sigma_\theta'\right)h^2 + O(h^3). \tag{80}$$

Adding the mean-difference and covariance contributions gives the stated $J_{\mathrm{Gauss}}(\theta)$. For a vector-valued condition, replacing h by a vector increment gives the Fisher information matrix

$$[J_{\mathrm{Gauss}}(\theta)]_{ij} = (\partial_i\mu_\theta)^\top\Sigma_\theta^{-1}(\partial_j\mu_\theta) + \frac{1}{2}\operatorname{tr}\big(\Sigma_\theta^{-1}(\partial_i\Sigma_\theta)\Sigma_\theta^{-1}(\partial_j\Sigma_\theta)\big). \tag{81}$$

□

Proposition A.8 (Averaging local affine coefficients gives linear Fisher information)**.** *Under the setup of Proposition A.7, suppose the fitted conditional affine velocity is*

$$v_\xi(x,t) = \dot\beta_t\mu_\xi + A_{\xi,t}(x - \beta_t\mu_\xi), \qquad \xi \in \{\theta, \theta+h\}, \tag{82}$$

and assume $A_{\xi,t}$ is jointly continuous in (ξ,t) and continuously differentiable in ξ on a neighborhood of θ times $[0,1]$. Average the two fitted linear coefficients,

$$\bar A_{\theta,h,t} = \frac{1}{2}\left(A_{\theta,t} + A_{\theta+h,t}\right), \tag{83}$$

and construct two new velocities

$$\widetilde v_\theta^{(h)}(x,t) = \dot\beta_t\mu_\theta + \bar A_{\theta,h,t}(x - \beta_t\mu_\theta), \tag{84}$$

$$\widetilde v_{\theta+h}^{(h)}(x,t) = \dot\beta_t\mu_{\theta+h} + \bar A_{\theta,h,t}(x - \beta_t\mu_{\theta+h}). \tag{85}$$

Let $\widetilde q_\theta^{(h)}$ and $\widetilde q_{\theta+h}^{(h)}$ denote their endpoint distributions from the common source $\rho_0 = \mathcal{N}(0,I)$. Then

$$D_{\mathrm{J}}\left(\widetilde q_\theta^{(h)}, \widetilde q_{\theta+h}^{(h)}\right) = \mu_\theta'^{\top}\Sigma_\theta^{-1}\mu_\theta'\,h^2 + O(h^3), \tag{86}$$

and therefore

$$\lim_{h\to 0}\frac{D_{\mathrm{J}}\left(\widetilde q_\theta^{(h)}, \widetilde q_{\theta+h}^{(h)}\right)}{h^2} = \mu_\theta'^{\top}\Sigma_\theta^{-1}\mu_\theta', \tag{87}$$

which is linear Fisher information.

Proof. Applying Lemma A.1 to the two constructed velocities shows that their endpoint distributions are Gaussian. Because the velocities have the same linear coefficient $\bar A_{\theta,h,t}$ and start from the same Gaussian source, their endpoint covariance is identical; denote it by $\bar\Sigma_{\theta,h}$. Their respective endpoint means are μ_θ and $\mu_{\theta+h}$, so

$$\widetilde q_\theta^{(h)} = \mathcal{N}(\mu_\theta, \bar\Sigma_{\theta,h}), \qquad \widetilde q_{\theta+h}^{(h)} = \mathcal{N}(\mu_{\theta+h}, \bar\Sigma_{\theta,h}). \tag{88}$$

For Gaussians with a common covariance, Jeffreys divergence is exactly

$$D_{\mathrm{J}}\left(\widetilde q_\theta^{(h)}, \widetilde q_{\theta+h}^{(h)}\right) = (\mu_{\theta+h} - \mu_\theta)^\top\bar\Sigma_{\theta,h}^{-1}(\mu_{\theta+h} - \mu_\theta). \tag{89}$$

Since $\mu_{\theta+h} - \mu_\theta = \mu_\theta' h + O(h^2)$, Eq. (89) has the claimed expansion provided that

$$\bar\Sigma_{\theta,h} = \Sigma_\theta + O(h), \tag{90}$$

because matrix inversion is smooth in a neighborhood of the positive-definite matrix Σ_θ. Thus, it remains only to prove Eq. (90).

Because $A_{\theta+h,t} = A_{\theta,t} + O(h)$ uniformly on $t \in [0,1]$, we have $\bar A_{\theta,h,t} = A_{\theta,t} + O(h)$. Let $\bar Q_t$ and $Q_{\theta,t}$ be the covariance trajectories generated by $\bar A_{\theta,h,t}$ and $A_{\theta,t}$, respectively. Both satisfy the affine

covariance ODE in Eq. (69), with the corresponding linear coefficient. Define $E_t = \bar{Q}_t - Q_{\theta,t}$ and $\Delta A_t = \bar{A}_{\theta,h,t} - A_{\theta,t}$. Subtracting these two instances of Eq. (69) gives

$$\dot{E}_t = \bar{A}_{\theta,h,t} E_t + E_t \bar{A}_{\theta,h,t}^\top + \Delta A_t Q_{\theta,t} + Q_{\theta,t} \Delta A_t^\top, \qquad E_0 = 0. \tag{91}$$

The coefficients and $Q_{\theta,t}$ are bounded on $t \in [0,1]$, while $\sup_t \|\Delta A_t\| = O(h)$. Therefore Eq. (91) implies $\|E_t\| \le C_1 \int_0^t \|E_s\|\, ds + C_2 h$, and Gronwall's inequality gives $\sup_{t\in[0,1]} \|E_t\| = O(h)$. At $t = 1$, $\bar{Q}_1 = \bar{\Sigma}_{\theta,h}$ and $Q_{\theta,1} = \Sigma_\theta$, so

$$\bar{\Sigma}_{\theta,h} - \Sigma_\theta = E_1 = O(h), \tag{92}$$

which proves Eq. (90). Positive definiteness of Σ_θ then gives $\bar{\Sigma}_{\theta,h}^{-1} = \Sigma_\theta^{-1} + O(h)$. Substituting this relation and $\mu_{\theta+h} - \mu_\theta = \mu'_\theta h + O(h^2)$ into Eq. (89) yields

$$D_{\mathrm{J}}\left(\widetilde{q}_\theta^{(h)}, \widetilde{q}_{\theta+h}^{(h)}\right) = \mu'_\theta{}^\top \Sigma_\theta^{-1} \mu'_\theta\, h^2 + O(h^3), \tag{93}$$

□

Proposition A.9 (Unconstrained flows give Jeffreys divergence and full Fisher information)**.** *Consider an unconstrained flow trained by flow matching. At the population optimum, the standard flow-matching result (Lipman et al., 2023) gives*

$$q_c^{\mathrm{uncon}} = p_c. \tag{94}$$

Hence the induced distance is the true Jeffreys divergence:

$$D_{\mathrm{uncon}}(a, b) = D_{\mathrm{J}}(p_a, p_b). \tag{95}$$

For a continuous parameter $\theta \in \mathbb{R}^k$, further assume p_θ is positive on a common support, $\log p_\theta(x)$ is three times continuously differentiable in θ, and differentiation can be exchanged with integration. Then, for $h \in \mathbb{R}^k$,

$$D_{\mathrm{J}}(p_\theta, p_{\theta+h}) = h^\top J(\theta) h + o(\|h\|^2), \tag{96}$$

where $J(\theta)$ is the full Fisher information matrix.

Proof. The population endpoint identity $q_c^{\mathrm{uncon}} = p_c$ is the standard flow-matching result stated above and is not reproved here. Substitution into Eq. (6) gives $D_{\mathrm{uncon}}(a, b) = D_{\mathrm{J}}(p_a, p_b)$.

For the continuous family, define the score $s_i(x;\theta) = \partial_{\theta_i} \log p_\theta(x)$ and the Fisher information matrix

$$J_{ij}(\theta) = \mathbb{E}_{p_\theta}[s_i(X;\theta) s_j(X;\theta)]. \tag{97}$$

The stated regularity conditions imply $\mathbb{E}_{p_\theta}[s_i(X;\theta)] = 0$ and the information identity

$$J_{ij}(\theta) = -\mathbb{E}_{p_\theta}[\partial_{\theta_i} \partial_{\theta_j} \log p_\theta(X)]. \tag{98}$$

Expand $\ell_{\theta+h}(x) = \log p_{\theta+h}(x)$ around θ:

$$\ell_{\theta+h}(x) = \ell_\theta(x) + h^\top \nabla_\theta \ell_\theta(x) + \frac{1}{2} h^\top \nabla_\theta^2 \ell_\theta(x) h + o(\|h\|^2). \tag{99}$$

Taking expectation under p_θ yields

$$D_{\mathrm{KL}}(p_\theta \| p_{\theta+h}) = \frac{1}{2} h^\top J(\theta) h + o(\|h\|^2). \tag{100}$$

Applying the same expansion at $\theta + h$ with increment $-h$ gives

$$D_{\mathrm{KL}}(p_{\theta+h} \| p_\theta) = \frac{1}{2} h^\top J(\theta + h) h + o(\|h\|^2) = \frac{1}{2} h^\top J(\theta) h + o(\|h\|^2), \tag{101}$$

where the last equality uses continuity of J. Adding both directions yields

$$D_{\mathrm{J}}(p_\theta, p_{\theta+h}) = h^\top J(\theta) h + o(\|h\|^2). \tag{102}$$

□

Proposition A.7 should not be interpreted as a special case of Proposition A.9. The former constrains the velocity to be affine and can be applied even when the target distributions are non-Gaussian. The latter instead assumes an unconstrained flow that recovers the target distributions themselves.

Proposition A.10 (Similarity-constrained velocities generate similarity transformations)**.** *Let T be the time-one map of*

$$\frac{dx}{dt} = a_t + \Omega_t(x - o_t) + \lambda_t(x - o_t), \qquad \Omega_t^\top = -\Omega_t. \tag{103}$$

If the coefficients are continuous on $[0, 1]$, then

$$T(z) = sRz + u, \qquad s > 0, \quad R \in \mathrm{SO}(d), \quad u \in \mathbb{R}^d. \tag{104}$$

Thus the flow transforms every data point by an isotropic scaling, a rotation, and a translation, without shear or nonlinear deformation.

Proof. Rewrite the flow as $\dot{x} = A_t x + r_t$, where

$$A_t = \Omega_t + \lambda_t I, \qquad r_t = a_t - (\Omega_t + \lambda_t I)o_t. \tag{105}$$

Let Φ_t solve $\dot{\Phi}_t = A_t\Phi_t$ with $\Phi_0 = I$. Since Ω_t is skew-symmetric,

$$\frac{d}{dt}(\Phi_t^\top \Phi_t) = \Phi_t^\top (A_t^\top + A_t)\Phi_t \tag{106}$$

$$= 2\lambda_t \Phi_t^\top \Phi_t. \tag{107}$$

Let $M_t = \Phi_t^\top \Phi_t$. Because $M_0 = I$, the matrix ODE $\dot{M}_t = 2\lambda_t M_t$ has the solution

$$M_t = \exp\left(2\int_0^t \lambda_\tau \, d\tau\right) I = s_t^2 I, \qquad s_t = \exp\left(\int_0^t \lambda_\tau \, d\tau\right) > 0. \tag{108}$$

Define $R_t = s_t^{-1}\Phi_t$. Then

$$R_t^\top R_t = s_t^{-2}\Phi_t^\top \Phi_t = I, \tag{109}$$

so R_t is orthogonal. Moreover, $R_0 = I$ and R_t varies continuously with t. Since an orthogonal matrix has determinant $+1$ or -1, continuity prevents the determinant from changing sign; hence $\det R_t = 1$ and $R_t \in \mathrm{SO}(d)$.

Finally, variation of constants gives the trajectory starting from z as

$$T_t(z) = \Phi_t z + \Phi_t \int_0^t \Phi_\tau^{-1} r_\tau \, d\tau = s_t R_t z + u_t, \tag{110}$$

where $u_t = \Phi_t \int_0^t \Phi_\tau^{-1} r_\tau \, d\tau$ is independent of z. Evaluating at $t = 1$ and writing $s = s_1$, $R = R_1$, and $u = u_1$ yields $T(z) = sRz + u$, as claimed. □

B Likelihood Fine-Tuning and Metric Correspondences

B.1 Optional negative-log-likelihood fine-tuning

Evaluating the Jeffreys divergence requires log densities of $q_c^{\mathcal{V}}$. Flow matching optimizes a regression objective that is fast to train but does not directly maximize likelihood. Many previous works emploied negative log likelihood (NLL) to improve fitting. In this manuscript, we applied fine-tuning for our defined geometric-preserved distance (Supplementary Figure S12).

Below we describe the fine-tuning procedure, and later prove that optimizing NLL also recovers the distance correspondences in Table 1 at their population optima.

The log density is evaluated using the continuous change-of-variables identity

$$\log q_c^{\mathcal{V}}(x_1) = \log \rho_0(x_0) - \int_0^1 \nabla_x \cdot v_c^{\mathcal{V}}(x_t, t) \, dt, \qquad x_0 = (T_{c,1}^{\mathcal{V}})^{-1}(x_1). \tag{111}$$

Here, $\nabla_x \cdot v_c^{\mathcal{V}}$ measures how the flow locally expands or contracts volume. For an observed response x_1, we integrate backward to recover x_0 and evaluate $\log q_c^{\mathcal{V}}(x_1)$. We fine-tune the constrained velocity field by minimizing the NLL

$$\mathcal{L}_{\mathrm{NLL}} = -\frac{1}{N}\sum_c \sum_{i=1}^{n_c} \log q_c^{\mathcal{V}}\left(x_i^{(c)}\right), \qquad N = \sum_c n_c. \tag{112}$$

This stage encourages $q_c^{\mathcal{V}}$ to assign higher density to the recorded responses, improving its finite-sample approximation to p_c while keeping the velocity constrained to $\mathcal{V}$.

B.2 Likelihood-based proofs

This section proves the counterparts of Table 1 when $q_c^{\mathcal{V}}$ is fitted by population negative log likelihood. These statements describe the global optimum of Eq. (112). By Lemma A.1, affine velocity classes transport the Gaussian source within a Gaussian family. We assume finite second moments and positive-definite covariances wherever an affine Gaussian density is used.

For the translation families below, let $\mathbf{b} = (b_c)_{c\in\mathcal{C}}$ and define the population objective

$$\begin{aligned}\mathcal{L}_{\mathrm{NLL}}(\mathbf{b}) &= \mathbb{E}_{C\sim\pi_{\mathcal{C}}}\mathbb{E}_{X\sim p_C}[-\log q_C^{\mathcal{V}}(X)] \\ &= \sum_{c\in\mathcal{C}} \pi_{\mathcal{C}}(c)\,\mathcal{L}_{\mathrm{NLL},c}(b_c), \\ \mathcal{L}_{\mathrm{NLL},c}(b_c) &= -\mathbb{E}_{X\sim p_c}[\log q_c^{\mathcal{V}}(X)].\end{aligned} \tag{113}$$

Proposition B.1 (Likelihood-trained translations give Euclidean, cosine, and correlation distances)**.** *Let $\rho_0 = \mathcal{N}(0, I)$ and $v_c(x,t) = \dot\beta_t b_c$, with $\beta_0 = 0$ and $\beta_1 = 1$. Then the time-one map is $T_{c,1}(x) = x + b_c$, and Lemma A.1 gives $q_c^{\mathcal{V}} = \mathcal{N}(b_c, I)$. Assume $\pi_{\mathcal{C}}(c) > 0$ and that each p_c has finite second moments. Let $r > 0$ and $H = I - \mathbf{1}\mathbf{1}^\top/d$. Population likelihood training yields:*

$$b_c^\star = \mu_c \qquad \textit{without an additional constraint}, \tag{114}$$

$$b_c^\star = r\frac{\mu_c}{\|\mu_c\|} \qquad \textit{under } \|b_c\| = r, \quad \mu_c \neq 0, \tag{115}$$

$$b_c^\star = r\frac{H\mu_c}{\|H\mu_c\|} \qquad \textit{under } \|b_c\| = r,\ \mathbf{1}^\top b_c = 0, \quad H\mu_c \neq 0. \tag{116}$$

The corresponding Jeffreys divergences are squared Euclidean distance, $2r^2$ times cosine distance, and $2r^2$ times Pearson correlation distance, respectively.

Proof. Equation (113) shows that the global likelihood objective is a positive weighted sum of condition-specific losses. Because b_c is condition specific, its minimization separates over c. Direct evaluation of the Gaussian negative log likelihood gives

$$\mathcal{L}_{\mathrm{NLL},c}(b_c) = \frac{1}{2}\mathbb{E}_{X\sim p_c}\left[\|X - b_c\|_2^2\right] + \mathrm{const} \tag{117}$$

$$= \frac{1}{2}\|b_c - \mu_c\|_2^2 + \frac{1}{2}\operatorname{tr}(\Sigma_c) + \mathrm{const} \tag{118}$$

$$= \frac{1}{2}\|b_c\|^2 - b_c^\top \mu_c + \mathrm{const}, \tag{119}$$

where terms independent of b_c are absorbed into the constant. The unconstrained quadratic $\frac{1}{2}\|b_c - \mu_c\|_2^2$ is uniquely minimized at $b_c = \mu_c$, proving Eq. (114).

On the sphere $\|b_c\| = r$, the norm term is constant, so Cauchy–Schwarz reduces the problem to maximizing $b_c^\top \mu_c$ and gives Eq. (115). On the centered sphere, $b_c = Hb_c$, so $b_c^\top \mu_c = b_c^\top H\mu_c$; applying Cauchy–Schwarz within this subspace gives Eq. (116).

All three fitted distributions have identity covariance, so Jeffreys divergence equals the squared distance between their means. Substitution gives

$$D_{\mathcal{V}_{\mathrm{Euc}}}(a,b) = \|\mu_a - \mu_b\|^2, \tag{120}$$

$$D_{\mathcal{V}_{\cos}}(a,b) = 2r^2\left(1 - \frac{\mu_a^\top \mu_b}{\|\mu_a\|\|\mu_b\|}\right), \tag{121}$$

$$D_{\mathcal{V}_{\mathrm{corr}}}(a,b) = 2r^2\left(1 - \frac{(H\mu_a)^\top (H\mu_b)}{\|H\mu_a\|\|H\mu_b\|}\right). \tag{122}$$

If a required mean vector is zero, the corresponding constrained optimum is nonunique, exactly matching the usual domain restriction of cosine or correlation distance. The response distributions p_c need not be Gaussian. □

Proposition B.2 (Likelihood-trained affine flows give Mahalanobis and Gaussian Jeffreys divergence)**.** *Assume $\pi_{\mathcal{C}}(c) > 0$ and each p_c has mean μ_c and positive-definite covariance Σ_c. Let $\Sigma_{\mathrm{pool}} = \mathbb{E}_{C\sim\pi_{\mathcal{C}}}[\Sigma_C]$.*

If the reachable Gaussian family has condition-specific means and one shared covariance,

$$q_c^{\mathcal{V}} = \mathcal{N}(m_c, S), \tag{123}$$

then the population likelihood optimum is $m_c^\star = \mu_c$, $S^\star = \Sigma_{\mathrm{pool}}$, and

$$D_{\mathcal{V}}(a,b) = (\mu_a - \mu_b)^\top \Sigma_{\mathrm{pool}}^{-1}(\mu_a - \mu_b). \tag{124}$$

If both means and covariances are condition specific, the optimum is $q_c^{\mathcal{V}} = \mathcal{N}(\mu_c, \Sigma_c)$ and the induced distance is the Gaussian Jeffreys divergence in Proposition A.6.

Proof. For the shared-covariance family, direct evaluation of the Gaussian negative log likelihood and averaging over $C \sim \pi_{\mathcal{C}}$ gives

$$\mathcal{L}_{\mathrm{NLL}}(\mathbf{m}, S) = \frac{1}{2}\left[\log\det S + \mathbb{E}_C \operatorname{tr}(S^{-1}\Sigma_C) + \mathbb{E}_C(\mu_C - m_C)^\top S^{-1}(\mu_C - m_C)\right] + \text{const.} \tag{125}$$

For every fixed $S \succ 0$, positivity of $\pi_{\mathcal{C}}(c)$ and positive definiteness of S^{-1} give the unique mean optimum $m_c^\star = \mu_c$. The remaining covariance objective is

$$\frac{1}{2}\left[\log\det S + \operatorname{tr}(S^{-1}\Sigma_{\mathrm{pool}})\right] + \text{const.} \tag{126}$$

The Gaussian cross-entropy is uniquely minimized by matching the pooled covariance, so $S^\star = \Sigma_{\mathrm{pool}}$. Hence

$$q_c^{\mathcal{V}} = \mathcal{N}(\mu_c, \Sigma_{\mathrm{pool}}), \qquad D_{\mathcal{V}}(a,b) = (\mu_a - \mu_b)^\top \Sigma_{\mathrm{pool}}^{-1}(\mu_a - \mu_b), \tag{127}$$

where the distance follows from the common-covariance Gaussian KL formula.

If the covariance is condition specific, the global loss separates over c. Minimizing the Gaussian cross-entropy conditionwise gives $(m_c^\star, S_c^\star) = (\mu_c, \Sigma_c)$ for every condition. Adding the two directed Gaussian KL divergences cancels the log determinants and yields the formula in Proposition A.6. □

Proposition B.3 (Likelihood-trained affine flows give Gaussian and linear Fisher information)**.** *Let θ be scalar, and suppose likelihood optimization is pointwise-separable in θ and can realize $q_\theta^{\mathcal{V}} = \mathcal{N}(m_\theta, S_\theta)$ at every condition. Assume μ_θ and Σ_θ are twice continuously differentiable and Σ_θ is positive definite. At the population likelihood optimum, $m_\theta^\star = \mu_\theta$ and $S_\theta^\star = \Sigma_\theta$, and*

$$\lim_{h\to 0} \frac{D_{\mathcal{V}}(\theta, \theta+h)}{h^2} = \mu_\theta'^\top \Sigma_\theta^{-1}\mu_\theta' + \frac{1}{2}\operatorname{tr}\left(\Sigma_\theta^{-1}\Sigma_\theta'\Sigma_\theta^{-1}\Sigma_\theta'\right). \tag{128}$$

If the two nearby fitted distributions are instead constrained to share a covariance $\bar{\Sigma}_{\theta,h} = \Sigma_\theta + O(h)$ while retaining their separate means, the limit is $\mu_\theta'^\top \Sigma_\theta^{-1}\mu_\theta'$, the linear Fisher information.

Proof. At each fixed θ, minimizing the Gaussian cross-entropy gives the unique pointwise optimum

$$q_\theta^{\mathcal{V}} = \mathcal{N}(\mu_\theta, \Sigma_\theta). \tag{129}$$

Proposition A.7 then applies directly and gives

$$D_{\mathcal{V}}(\theta, \theta + h) = \left[\mu_\theta'^{\top} \Sigma_\theta^{-1} \mu_\theta' + \frac{1}{2} \operatorname{tr}\left(\Sigma_\theta^{-1} \Sigma_\theta' \Sigma_\theta^{-1} \Sigma_\theta' \right) \right] h^2 + O(h^3). \tag{130}$$

Under the shared-local-covariance constraint, Proposition A.8 gives

$$D_{\mathcal{V}}(\theta, \theta + h) = \mu_\theta'^{\top} \Sigma_\theta^{-1} \mu_\theta' \, h^2 + O(h^3). \tag{131}$$

Dividing by h^2 and taking $h \to 0$ proves both limits. The pointwise-realizability assumption is essential: if a finite neural parameterization couples distant values of θ, its population optimum need not equal the separate Gaussian moment projection at every stimulus. □

C A NEW DISTANCE METRIC THAT PRESERVES THE GEOMETRY OF A TEMPLATE

A framework for distance metrics makes it possible to design new metrics in a principled way. A design consists of two choices: a source distribution, which specifies the basic structure used to approximate each response distribution; and a velocity class, which specifies the permitted transformations from source to target.

As a demonstration, we introduce a metric that estimates the distance between two distributional approximations constrained to preserve the geometry of a prescribed template. Such a metric can be useful when the geometry of the target distribution is known a priori, or when one wishes to emphasize a particular geometric property of the distribution.

Geometric-template source distribution. Let $\mathcal{U} \subseteq \mathbb{R}^m$ be a bounded parameter domain, typically with $m < d$, and let $f : \mathcal{U} \to \mathbb{R}^d$ be an a priori specified template embedding. Thus, $u \in \mathcal{U}$ indexes a location on the template; for example, $\mathcal{U} = [0, 2\pi)$ and $f(u) = (\cos u, \sin u)$ trace a circle. Define the condition-indexed source distribution $\rho_{\mathrm{geom},c}$ by

$$Z_c = f(U) + \varepsilon_c, \qquad U \sim \mathrm{Unif}(\mathcal{U}), \qquad \varepsilon_c \sim \mathcal{N}(0, \sigma_c^2 I). \tag{132}$$

The additive Gaussian noise makes the source full dimensional, so it has a well-defined density in $\mathbb{R}^d$, while σ_c controls the thickness of the distribution around the template. The template geometry f is fixed across conditions; only its admissible transformation and its noise thickness are fitted.

Similarity-constrained velocity field. We want each fitted transformation to preserve the template's shape while allowing its position, orientation, and overall scale to vary. We therefore restrict the time-one map Eq. (2) to the orientation-preserving similarity group (Hall, 2015),

$$T_c(z) = s_c R_c z + t_c, \qquad s_c > 0, \quad R_c \in \mathrm{SO}(d). \tag{133}$$

Here t_c translates the template, R_c rotates it, and s_c changes its scale uniformly in every direction. These similarity transformations exclude reflection, shear, and nonlinear deformation.

To parameterize these transformations continuously, we constrain the velocity to the Lie algebra of this group:

$$\frac{dx}{dt} = a_{c,t} + \Omega_{c,t}(x - o_{c,t}) + \lambda_{c,t}(x - o_{c,t}), \qquad \Omega_{c,t}^{\top} = -\Omega_{c,t}. \tag{134}$$

The quantities $a_{c,t}$, $\Omega_{c,t}$, $o_{c,t}$, and $\lambda_{c,t}$ are learnable parameters produced by a neural network. The vector $a_{c,t}$ is the infinitesimal generator of translation, the matrix $\Omega_{c,t}$ generates rotation about $o_{c,t}$, and the scalar $\lambda_{c,t}$ generates isotropic expansion or contraction. Integrating these infinitesimal generators over time produces a similarity transformation of the form above, as proved in Proposition A.10.

Training and induced geometry-template dissimilarity. Training proceeds in two stages. First, we fix each noise scale σ_c at its initial value and optimize the neural-network parameters of the similarity-constrained velocity field using the flow-matching objective in Eq. (5). Second, we make the positive scale σ_c learnable and jointly fine-tune σ_c and the velocity parameters by minimizing the NLL objective in Eq. (112). This stage adapts the template noiseless to the observations. After training, we define the geometry-template dissimilarity between conditions a and b as their Jeffreys divergence as before.

C.1 Demonstration of a new distance metric that preserves the geometry of a template

Besides estimation, having a framework allows one to design new distance metrics. In this section, we demonstrate the geometry-preserving metric from Section C on a toy dataset. This metric fits each target by a distribution that preserves the geometry of a chosen source template.

The synthetic dataset contains two conditions, each represented by a noisy square with a different orientation. The source distribution is a unit square with a small amount of noise. We fit each condition using either a similarity-constrained velocity field, which permits only translation, rotation, and isotropic scaling of the source distribution, or an unconstrained nonlinear velocity field. The similarity-constrained model better recovers the ground truth and preserves the square geometry, whereas the unconstrained model distorts it (Figure S12). This combination of a geometry-informed source template and a similarity-constrained velocity field is therefore useful when the data geometry is known or when a particular geometric structure should be preserved.

D Experimental Details

We organize the experimental details into four groups: estimator specifications, simulation datasets, real neural datasets, and the joint estimator–dataset specification for the geometry-template Jeffreys distance.

D.1 Estimators

D.1.1 Flow matching training

For affine flow which contains a velocity with A matrix, for example, $v_c(x,t) = \dot{\beta}_t b_c + A_{c,t}(x - \beta_t b_c)$, the symmetric matrix $A_{c,t}$ contains $d(d+1)/2 = O(d^2)$ learned entries. This parameter growth can make high-dimensional fits unstable with finite samples, so we use the regularized loss:

$$\mathcal{L}_\lambda = \mathcal{L}_{\mathrm{FM}} + \lambda\, \mathbb{E}_{c,t}\left[\frac{1}{d}\|A_{c,t}\|_F^2\right].$$

We compare six regularization strengths, $\lambda \in \{0, 0.005, 0.05, 0.5, 5, 50\}$. We first fit an unregularized model and save its checkpoint. Starting from this shared checkpoint, we then continue fitting separately for each value of λ. All resulting candidate checkpoints are evaluated on the same validation set, and we select the candidate with the minimum unpenalized validation Flow Matching loss (Section D.1.2).

For unconstrained velocity, we don't use regularization. We instead divide the fitting data into training and validation sets, apply early stopping with a patience of 1000 epochs, and retain the checkpoint with the lowest validation loss.

Additional training details. Unless otherwise specified, source samples X_0 are drawn independently from $\mathcal{N}(0, I_d)$ and independently of the observed responses. At each optimization step, we pair source samples with observed responses, draw the flow time uniformly as $t \sim \mathrm{Unif}[5 \times 10^{-4}, 1 - 5 \times 10^{-4}]$, and minimize the loss along a cosine probability path. We optimize using AdamW (Loshchilov & Hutter, 2019) and clip the gradient norm at 10. The network architecture, batch size, and learning rate are experiment-specific and summarized in Table 2.

D.1.2 VELOCITY NETWORK ARCHITECTURE

Condition embedding. Raw conditions are converted to fixed feature vectors before they enter the velocity model. A categorical condition $c \in \{0, \dots, K-1\}$ is represented by the one-hot vector $\phi_{\mathrm{cat}}(c) = e_c \in \mathbb{R}^K$.

Throughout this subsection, $J = 8$ denotes the number of Gaussian radial-basis centers used for both nonperiodic and periodic scalar embeddings, as detailed below.

For a nonperiodic scalar condition s, let $\bar{s}_{\mathrm{tr}}$ and σ_{tr} denote the mean and root-mean-square scale of the training conditions. The standardized coordinate $z(s)$ and the linear-plus-RBF embedding are

$$\begin{aligned} z(s) &= \frac{s - \bar{s}_{\mathrm{tr}}}{\sigma_{\mathrm{tr}}}, \\ m_j &= z_{\min} + \frac{j-1}{J-1}(z_{\max} - z_{\min}), \qquad j = 1, \dots, J, \\ h &= \frac{z_{\max} - z_{\min}}{J-1}, \qquad \psi_j(s) = \exp\left[-\frac{1}{2}\left(\frac{z(s) - m_j}{h}\right)^2\right], \\ \phi_{\mathrm{np}}(s) &= \big(z(s), \psi_1(s), \dots, \psi_J(s)\big) \in \mathbb{R}^{J+1}, \qquad J = 8. \end{aligned}$$

Here $z_{\min}$ and $z_{\max}$ are the extrema of the standardized training conditions. Each Gaussian feature is largest near its center m_j and decays smoothly with distance, while retaining $z(s)$ preserves the global linear ordering. The continuous Fisher benchmarks therefore use a nine-dimensional condition vector.

For a period-P scalar, centers instead lie uniformly on a circle. With origin o, the periodic embedding uses the shortest wrapped displacement δ_P:

$$\begin{aligned} m_j &= o + \frac{j-1}{J}P, \qquad h = \frac{P}{J}, \qquad j = 1, \dots, J, \\ \delta_P(s, m_j) &= \mathrm{mod}\left(s - m_j + \frac{P}{2}, P\right) - \frac{P}{2}, \\ \psi_j^{\mathrm{per}}(s) &= \exp\left[-\frac{1}{2}\left(\frac{\delta_P(s, m_j)}{h}\right)^2\right], \\ \phi_{\mathrm{per}}(s) &= \big(\psi_1^{\mathrm{per}}(s), \dots, \psi_J^{\mathrm{per}}(s)\big) \in \mathbb{R}^J, \qquad J = 8. \end{aligned}$$

Consequently $\phi_{\mathrm{per}}(s + P) = \phi_{\mathrm{per}}(s)$, so nearby orientations across the periodic boundary remain nearby in feature space.

In the Stringer experiment, both affine and unconstrained flow use the eight-dimensional periodic embedding $\phi_{\mathrm{per}}(s)$ with $P = \pi$. For a mixed condition consisting of a categorical variable c and a scalar s, which may be periodic or non-periodic, we concatenate the two representations as $\phi_{\mathrm{mix}}(c, s) = [e_c, \phi(s)]$. We use this representation for the time-resolved OU experiments, where c denotes category and s denotes physical time. For both the Allen two-photon calcium-imaging and Neuropixels experiments, c indexes the 40 joint drifting-direction–temporal-frequency conditions, while s denotes physical time and is represented by $\phi_{\mathrm{np}}(s)$.

Network architecture. Responses are standardized using statistics fitted on the fitting responses. The scalar flow time is used directly without any embeddings. Writing $h_t(u) = [t, \phi(u)]$, where $\phi(u)$ denotes the appropriate categorical, scalar, periodic, or mixed embedding of the experimental condition described above. Neural architectures are feed-forward multilayer perceptrons with SiLU activation with the width and number of hidden layers listed in Table 2. Neural networks output velocity parameters:

$$\begin{aligned} b(c) &= \mathrm{MLP}_b(\phi(c)) \in \mathbb{R}^d \qquad \text{(endpoint mean)}, \\ a_{c,t} &= \mathrm{MLP}_A(h_t(c)) \in \mathbb{R}^{d(d+1)/2}, \qquad A_{c,t} = \mathrm{unpack}_{\mathrm{sym}}(a_{c,t}) = A_{c,t}^\top, \\ v_c^{\mathrm{aff}}(x_t, t) &= \dot{\beta}_t b(c) + A_{c,t}\big(x_t - \beta_t b(c)\big), \\ v_c^{\mathrm{uncon}}(x_t, t) &= \mathrm{MLP}_v([x_t, t, \phi(c)]) \in \mathbb{R}^d. \end{aligned}$$

Translation velocity sets $A_{c,t} = 0$, and shared affine replaces $A_{c,t}$ by a condition-independent A_t. The square-geometry models use the analogous parameterizations.

Table 2: Flow Matching training settings. Width × depth gives the hidden width and number of hidden layers.

Experiment	Flow family	Width × depth	Batch	Learning rate
Categorical RDM	constrained and nonlinear	128×3	3000	10^{-4} for correlation/cosine; 10^{-3} otherwise
Independent Gaussian Fisher	affine	256×3	256	10^{-4}
Independent mixture Fisher	affine and nonlinear	256×3	256	10^{-4}
Gaussian OU	affine	256×3	256	10^{-4}
Mixture OU	affine and nonlinear	256×3	256	10^{-4}
Stringer	affine and nonlinear	256×3	256	10^{-4}
Allen	affine and nonlinear	256×3	2048	10^{-4}
Square geometry	similarity-constrained and nonlinear	64×2	256	10^{-3} FM; 10^{-4} NLL

Jeffreys divergence estimation. When the velocity has the affine (linear-in-state) form $v_c^{\mathcal{V}}(x,t) = A_{c,t}x + r_{c,t}$, we integrate the mean and covariance ODEs in Eq. (14) and evaluate the corresponding Gaussian distance from the resulting moments; Gaussian Jeffreys divergence uses Eq. (16). For an unconstrained velocity—used to estimate Jeffreys divergence and full Fisher information—we instead generate endpoint samples by solving Eq. (1) and evaluate $\log q_c^{\mathcal{V}}(x)$ through the continuous change-of-variables identity in Eq. (111). Jeffreys divergence is then estimated with Eq. (8).

D.1.3 Gaussian-process with kernel regression

Our implementation follows the GKR code of Ye & Wessel (2025).

GKR models the smooth conditional response distribution as

$$q_{\mathrm{GKR}}(x \mid u) = \mathcal{N}\Big(x \mid \widehat{\mu}(u), \widehat{\Sigma}(u)\Big), \tag{135}$$

Here u denotes the conditioning input to each fitted GKR model: $u = \theta$ for the Fisher datasets, $u = (c,t)$ for the synthetic OU datasets, and u is grating orientation for Stringer. For the Allen two-photon calcium-imaging and Allen Neuropixels ephys datasets, we fit a separate GKR model for each joint drifting-direction–temporal-frequency condition, with $u = t$ within each model. We use an RBF kernel for nonperiodic inputs and a periodic kernel for grating orientation.

Let $r_i = x_i - \widehat{\mu}(u_i)$ be the fitted residuals. GKR estimates the conditional covariance by normalized kernel regression of their outer products,

$$\widehat{\Sigma}(u) = \frac{\sum_{i=1}^{n} k_\lambda(u_i, u) r_i r_i^\top}{\sum_{i=1}^{n} k_\lambda(u_i, u)} + \epsilon_\Sigma I. \tag{136}$$

For the nonperiodic scalar conditions, inputs are standardized and

$$k_\lambda(u_i, u) = \exp\left[-\frac{(\widetilde{u}_i - \widetilde{u})^2}{2(\lambda + \epsilon_\lambda)}\right]. \tag{137}$$

For an orientation with period P, the periodic covariance kernel is instead

$$k_\lambda^{\mathrm{per}}(u_i, u) = \exp\left[-\frac{\sin^2(\pi(u_i - u)/P)}{\lambda + \epsilon_\lambda}\right]. \tag{138}$$

The default configuration uses 200 inducing points; 300 mean-GP iterations at learning rate 0.05; and 30 covariance-kernel epochs at learning rate 0.1, with a validation fraction of $1/3$.

D.1.4 Wishart-process estimator

Our Wishart process implementation reuses the original JAX/NumPyro model, kernels, variational inference, and posterior-mode prediction from https://github.com/neurostatslab/wishart-process, pinned to commit `5607d14`.

The Wishart process models both the mean and covariance of a smoothly condition-dependent Gaussian distribution (Nejatbakhsh et al., 2023). For a conditioning input u and response $X \in \mathbb{R}^d$,

$$X \mid u \sim \mathcal{N}(\mu(u), \Sigma(u)), \qquad \mu_j(\cdot) \sim \mathrm{GP}(0, k_\mu), \qquad F_{jp}(\cdot) \sim \mathrm{GP}(0, k_\Sigma). \tag{139}$$

The latent functions μ_j model the conditional mean, while the factors F_{jp} model the conditional covariance; all mean coordinates and covariance factors have independent GP priors. Let $F(u) \in \mathbb{R}^{d\times d}$ collect the active covariance factors used by the upstream implementation, which defines

$$\Sigma(u) = L\left[F(u)F(u)^\top + \delta I_d\right] L^\top, \qquad \delta = 0.1. \tag{140}$$

The Gram matrix $F(u)F(u)^\top$ is positive semidefinite, and adding δI_d makes it positive definite before scaling by L. This Gram construction motivates the name *Wishart process*: at each condition, outer products of Gaussian latent factors form a covariance matrix, while the GP correlations across u make that covariance vary smoothly with the condition. The shared matrix L is initialized as the Cholesky factor of $(\widehat{\Sigma}_{\mathrm{pool}} + \delta I_d)/(d+1)$, where $\widehat{\Sigma}_{\mathrm{pool}}$ is the pooled within-bin empirical covariance. It is then optimized during SVI.

The mean and covariance-factor GPs use separate product kernels. For $r \in \{\mu, \Sigma\}$ and condition coordinate q,

$$k_r(u, u') = \prod_q \exp\left[-\frac{d_q(u_q, u'_q)^2}{\lambda_{r,q}}\right] + 10^{-3}\mathbf{1}\{u = u'\}, \tag{141}$$

$$d_q(a, b) = \begin{cases} a - b, & \text{for a nonperiodic coordinate,} \\ \sin(\pi(a-b)/P_q), & \text{for a coordinate with period } P_q. \end{cases} \tag{142}$$

Here $\lambda_{\mu,q}$ and $\lambda_{\Sigma,q}$ separately control how smoothly the mean and covariance factors vary along condition coordinate q: larger widths produce stronger sharing across nearby conditions. The Fisher datsets have one nonperiodic coordinate $u = \theta$, whereas the Stringer recordings have one periodic orientation coordinate $u = \theta$ with period $P = \pi$.

The authors' code interface requires an equal-trial tensor with shape $K \times C \times d$. After fixing the train, validation, and test splits, our wrapper bins only the training observations, sets $K = \min_c n_c$, and subsamples every occupied bin to K responses. All validation and test observations are retained. Mean and covariance kernel widths ($\lambda_{\mu,q}$ and $\lambda_{\Sigma,q}$) are selected by held-out Gaussian log likelihood.

Benchmark	Mean-kernel widths	Covariance-kernel widths
Linear Fisher	$\lambda_\mu \in \{5, 10, 20\}$	$\lambda_\Sigma \in \{20{,}480, 40{,}960, 81{,}920\}$
Gaussian-mixture full Fisher	$\lambda_\mu \in \{2.25, 9, 36\}$	$\lambda_\Sigma \in \{144, 576, 2304\}$
Stringer recordings (initial grid)	$\lambda_\mu \in \{0.00125 \cdot 2^j : j = 0, \ldots, 8\}$	$\lambda_\Sigma \in \{0.05 \cdot 2^j : j = 0, \ldots, 7\} \cup \{\infty\}$
Gaussian OU	$\lambda_{\mu,t} \in \{1, 4, 16, 64, 256\}$	$\lambda_{\Sigma,t} \in \{256 \cdot 4^j : j = 0, \ldots, 7\} \cup \{\infty\}$
Two-scale Gaussian-mixture OU	$\lambda_{\mu,t} \in \{1, 4, 16, 64, 256\}$	$\lambda_{\Sigma,t} \in \{16 \cdot 4^j : j = 0, \ldots, 8\} \cup \{\infty\}$
Allen Neuropixels PCA-10	$\lambda_\mu \in \{0.015625 \cdot 2^j : j = 0, \ldots, 6\} \cup \{4, \infty\}$	$\lambda_\Sigma \in \{0.25, 1, 4, 16, 64\} \cup \{1024, 16384, 262144, \infty\}$

Table 3: Kernel-width grids used for the Wishart-process method. An infinite covariance width denotes a covariance that is constant over the continuous variable (e.g. orientation or trial time) within each condition.

To search kernel widths, we evaluate the grids in Table 3, adaptively expanding the search space when needed. We continue the expansion until the validation curve exhibits an inverted-U shape, indicating a local optimum, or until the kernel width becomes sufficiently large to produce nearly constant smoothing. For example, supplementary Figure S5 shows the resulting search profiles for the Stringer recordings.

For the PCA-70 Allen Neuropixels analysis in Figure 7, we found it practically difficult to identify optimal kernel widths because validation often favored very large values, resulting in estimates

that were nearly constant over time. This behavior may reflect the high dimensionality and non-Gaussianity of the neural responses.

Therefore we instead reducing the responses to 10 principal components where the Wishart process produced more reasonable fits, as shown in Supplementary Figure S10. Affine flow matching remains best among Gaussian methods.

D.1.5 CROSS-FITTED OPTIMAL LINEAR ESTIMATOR (OLE)

We mainly follow Kanitscheider et al. (2015) for the locally optimal linear estimator and its finite-sample bias correction. OLE fits a linear decoder on a training set and projects held-out responses onto the resulting axis. Our reported achieved-information estimate applies the bias correction shown below and is clipped at zero.

Specifically, we estimate this quantity with five-fold cross-fitting independently for each adjacent pair $\theta_L < \theta_R$ on a 31-endpoint grid, where $h = \theta_R - \theta_L = 0.4$. Responses are assigned to disjoint radius-$h/2$ windows around the two endpoints. If either window contains fewer than eight responses, we instead use the nearest disjoint groups with eight responses per endpoint. For fold f, four folds estimate the endpoint means and Ledoit–Wolf covariances. We then form

$$\widehat{\mu}'_f = \frac{\widehat{\mu}_{R,f} - \widehat{\mu}_{L,f}}{h}, \qquad \widehat{\Sigma}_f = \frac{\widehat{\Sigma}_{L,f} + \widehat{\Sigma}_{R,f}}{2}, \tag{143}$$

$$\widehat{w}_f = \frac{\widehat{\Sigma}_f^{-1}\widehat{\mu}'_f}{\widehat{\mu}'_f{}^{\top}\widehat{\Sigma}_f^{-1}\widehat{\mu}'_f}, \qquad \widehat{b}_f = \frac{\theta_L + \theta_R}{2} - \widehat{w}_f^{\top}\frac{\widehat{\mu}_{L,f} + \widehat{\mu}_{R,f}}{2}. \tag{144}$$

The fitted decoder $\widehat{\theta} = \widehat{w}_f^{\top} X + \widehat{b}_f$ is frozen and applied only to the held-out fold. Pooling the five held-out folds therefore evaluates each selected response exactly once out of fold. Let $\bar{y}_L, \bar{y}_R, s_L^2, s_R^2$, and n_L, n_R denote the means, sample variances, and counts of the pooled held-out projections. The reported estimate is

$$\widehat{I}_{\mathrm{OLE}}^{\mathrm{raw}} = \frac{(\bar{y}_R - \bar{y}_L)^2 - s_L^2/n_L - s_R^2/n_R}{h^2(s_L^2 + s_R^2)/2}, \qquad \widehat{I}_{\mathrm{OLE}} = \max\left\{\widehat{I}_{\mathrm{OLE}}^{\mathrm{raw}}, 0\right\}. \tag{145}$$

D.1.6 TRE DENSITY-RATIO ESTIMATION FOR CATEGORICAL JEFFREYS DIVERGENCE

For the full Jeffreys divergence comparison in Supplementary Figure S1, we use telescoping density-ratio estimation (TRE) (Rhodes et al., 2020) with $B \in \{1, 8, 16\}$ bridges. For each unordered pair of categorical conditions a and b, TRE draws independent samples $X_a \sim p_a$ and $X_b \sim p_b$ and forms $B + 1$ angular waymarks

$$Z_m = \cos(\phi_m)X_a + \sin(\phi_m)X_b, \qquad \phi_m = \frac{m\pi}{2B}, \qquad m = 0, \ldots, B. \tag{146}$$

Then we constructed B adjacent-ratio classifiers share a three-hidden-layer MLP trunk with width 128 and use separate linear output heads. Each condition-pair model is trained for at most 1000 epochs with batch size 512, AdamW learning rate 10^{-3}, zero weight decay, and gradient clipping at norm 10. We select the checkpoint with the lowest validation loss. If q_m denotes the density of Z_m, classifier head m is trained with equal class priors to distinguish q_m from q_{m+1}. The estimated logits then telescope to

$$\widehat{r}_{ab}(x) = \sum_{m=0}^{B-1} \widehat{\ell}_{ab,m}(x) \approx \sum_{m=0}^{B-1} \log \frac{q_m(x)}{q_{m+1}(x)} = \log \frac{p_a(x)}{p_b(x)}. \tag{147}$$

We estimate the Jeffreys divergence for each condition pair by averaging this log ratio under the two endpoint distributions and clip negative finite-sample estimates at zero:

$$\widehat{D}_{\mathrm{J}}(p_a, p_b) = \max\left\{0, \frac{1}{n_a}\sum_{i=1}^{n_a} \widehat{r}_{ab}(X_{a,i}) - \frac{1}{n_b}\sum_{j=1}^{n_b} \widehat{r}_{ab}(X_{b,j})\right\}. \tag{148}$$

D.2 SIMULATION DATASETS

D.2.1 CATEGORICAL RDM DATASET

Here, we describe the categorical RDM experiment shown in Figure 3.

We define the conditional response distribution as

$$z \mid Y = k \sim \mathcal{N}(\mu_k, \mathrm{diag}(v_k)), \tag{149}$$

where Y is uniformly distributed over $\{0, \dots, 4\}$ and z is three-dimensional. We generate the parameters μ_k and $\mathrm{diag}(v_k)$ once and hold them fixed across all five repetitions. Thus, the repetitions share the same category-conditional distributions and differ only in their sampled observations.

Specifically, we generate μ_k and $\mathrm{diag}(v_k)$ as follows. We draw each mean coordinate as $\mu_{k,j} = G_{k,j} Z_{k,j}$, where $G_{k,j} \sim \mathrm{Unif}(0.2, 2.0)$ and $Z_{k,j} \sim \mathrm{Unif}(0, 1)$. We resample the candidate means until every pair is separated by at least $\sqrt{3}/2$ (half the diagonal of the unit cube), thereby keeping the components well separated. We set the conditional variances to

$$v_{k,j} = 3(0.15)^2 + 0.15|\mu_{k,j}| + 10^{-5}. \tag{150}$$

We then independently draw five datasets from this fixed population. When fitting the structured models—that is, all velocity families in Table 1 except the unconstrained family—we use all observations. Because the response dimension is low, we set $\lambda = 0$ for affine flow, corresponding to no regularization. The nonlinear Jeffreys model instead uses an 80/20 training–validation split. Figure 3A shows accuracy as a function of sample size, and Figure 3B shows the corresponding RDM estimates at 2,000 total samples.

Figure S1 compares flow matching against plug-in mean estimators for correlation, cosine, and squared Euclidean distance; pooled-covariance Mahalanobis estimators with and without Ledoit–Wolf (LW) shrinkage; a condition-specific LW covariance estimator for Gaussian Jeffreys divergence; and TRE estimators using 1, 8, or 16 bridges for full Jeffreys divergence.

We compute all six ground-truth RDMs analytically from the known population means μ_k and diagonal covariances $\mathrm{diag}(v_k)$.

We evaluate estimation performance using the mean relative absolute error (MRAE) over the unique off-diagonal category pairs. Let $\mathcal{P} = \{(i, j) : 0 \le i < j \le C - 1\}$ and $M = |\mathcal{P}| = \binom{C}{2}$. For an estimated RDM $\widehat{D}$ and its ground-truth RDM D,

$$\mathrm{MRE}(\widehat{D}, D) = \frac{1}{M} \sum_{(i,j)\in\mathcal{P}} \frac{\left|\widehat{D}_{ij} - D_{ij}\right|}{\max\{|D_{ij}|, 10^{-12}\}}. \tag{151}$$

With $C = 5$, the average includes $M = 10$ unique category pairs.

D.2.2 FISHER DATASETS

Here we describe the datasets used for estimating Fisher information in Figure 4A–F. We first specify the Gaussian dataset used for linear Fisher information, then the mixture dataset used for full Fisher information.

Gaussian responses and linear Fisher information. For the Gaussian dataset (Figure 4A–C), $\theta \in [-6, 6]$, while the response dimension d follows the sweep protocol below:

$$p^{\mathrm{G}}(x \mid \theta) = \mathcal{N}(x \mid \mu(\theta), \Sigma(\theta)). \tag{152}$$

The mean is a Gaussian bump with equal-spacing centers $\vartheta_j = -6 + 12(j-1)/(d-1)$

$$\mu_j(\theta) = a_j \exp\left[-0.2\,(\theta - \vartheta_j)^2\right], \tag{153}$$

where $a_j \overset{\mathrm{iid}}{\sim} \mathrm{Unif}(0.2, 2.0)$. and the condition-dependent diagonal covariance is

$$[\Sigma(\theta)]_{j\ell} = \delta_{j\ell}\left[d(0.07)^2 + \frac{d}{300}\,1.30|\mu_j(\theta)| + 10^{-8}\right]. \tag{154}$$

The $d/300$ factor preserves the relative covariance modulation across response dimensions.

The target is linear Fisher information, $\mu'(\theta)^\top \Sigma(\theta)^{-1}\mu'(\theta)$, estimated by affine flow (Section D.1.1 and Proposition A.7), GKR (Section D.1.3), the Wishart process (Section D.1.4), and five-fold cross-fitted OLE (Section D.1.5). The sample-size sweep fixes $d = 300$ and uses $N \in \{500, 1000, 3000, 5000, 10000\}$; the dimension sweep fixes $N = 1000$ and uses $d \in \{10, 20, 50, 100, 200, 300, 400\}$.

For the affine-flow readout, each adjacent pair (θ_k, θ_{k+1}), with midpoint $\overline{\theta}_k = (\theta_k + \theta_{k+1})/2$, shares a covariance obtained by averaging its two affine matrices at every flow time:

$$\overline{A}_{k,t} = \tfrac{1}{2}\left(A_{\theta_k,t} + A_{\theta_{k+1},t}\right), \qquad \dot{\overline{\Sigma}}_{k,t} = \overline{A}_{k,t}\overline{\Sigma}_{k,t} + \overline{\Sigma}_{k,t}\overline{A}_{k,t}^\top, \qquad \overline{\Sigma}_{k,0} = I_d, \tag{155}$$

$$\widehat{I}_{\mathrm{lin}}(\overline{\theta}_k) = \frac{\Delta b_k^\top \overline{\Sigma}_{k,1}^{-1} \Delta b_k}{(\theta_{k+1} - \theta_k)^2}, \qquad \Delta b_k = b(\theta_{k+1}) - b(\theta_k). \tag{156}$$

For estimated Fisher information $\widehat{I}(\theta_k)$ and ground truth $I(\theta_k)$ at K evaluation points, we report MRAE

$$\mathrm{MRAE}(\widehat{I}, I) = \frac{1}{K}\sum_{k=1}^{K} \frac{\left|\widehat{I}(\theta_k) - I(\theta_k)\right|}{\max\{|I(\theta_k)|, 10^{-12}\}}. \tag{157}$$

We additionally report the Pearson correlation between the estimated and ground-truth Fisher curves.

Mixture responses and full Fisher information. For the non-Gaussian dataset (Figure 4D–F), we retain the Gaussian dataset's $\mu(\theta)$ and $\Sigma(\theta)$, but introduce an equally likely latent scale branch S:

$$\Pr(S = -1) = \Pr(S = +1) = \tfrac{1}{2}, \qquad X \mid \theta, S \sim \mathcal{N}(\mu(\theta), [1 + S\rho(\theta)]\Sigma(\theta)). \tag{158}$$

where $\rho(\theta) = 0.5 + 0.4\sin\left[\frac{\pi}{6}(\theta + 6) + 2.20\right]$. Because $\mathbb{E}[S] = 0$, the marginal mean and covariance remain exactly $\mu(\theta)$ and $\Sigma(\theta)$; $\rho(\theta)$ therefore changes higher-order structure without changing either of the first two moments, keeping the linear Fisher information the same. The target for the common error panels is the exact mixture full Fisher information,

$$I_{\mathrm{full}}(\theta) = \mathbb{E}_{X \sim p^{\mathrm{mix}}(\cdot\mid\theta)}\left[\left(\partial_\theta \log p^{\mathrm{mix}}(X \mid \theta)\right)^2\right]. \tag{159}$$

We use spacing $\Delta\theta = 0.4$ and compare Unconstrained Flow, Affine Flow, GKR, and Wishart. The sample-size sweep fixes $d = 20$ and uses $N \in \{1000, 4000, 7000, 10000, 13000\}$; the dimension sweep fixes $N = 7000$ and uses $d \in \{5, 20, 50, 80, 110\}$.

D.2.3 Time-resolved Jeffreys datasets

Here we describe the Gaussian and non-Gaussian OU datasets in Figure 6A–F.

Gaussian OU dataset: time-local marginals. The time-resolved dataset uses two conditions, $c \in \{0, 1\}$, response dimension d, and $T = 24$ time-bin centers $t_k = -6 + \left(k - \frac{1}{2}\right)\Delta t$, $\Delta t = 0.5$, $k = 1, \ldots, T$. For each tested d, the coordinatewise tuning-curve amplitudes are drawn once as $a_j \overset{\mathrm{iid}}{\sim} \mathrm{Unif}(0.2, 2.0)$, and the preferred time points are evenly spaced over the stimulus interval, $\vartheta_j = -6 + \frac{12(j-1)}{d-1}$, $j = 1, \ldots, d$. The common unsigned time-tuning curve is therefore

$$\mu_j(t) = a_j \exp\left[-0.2\,(t - \vartheta_j)^2\right], \qquad \mu(t) = \left(\mu_1(t), \ldots, \mu_d(t)\right)^\top. \tag{160}$$

Let $s_0 = 1$ and $s_1 = -1$. At any fixed time t, the marginal distribution is:

$$p_{\mathrm{G}}(x \mid c, t) = \mathcal{N}(x \mid s_c\, 0.7\mu(t), \Sigma_{\mathrm{G}}(t)), \tag{161}$$

$$[\Sigma_{\mathrm{G}}(t)]_{j\ell} = \delta_{j\ell}\left[d(0.07)^2 + \frac{d}{300}\,1.30|\mu_j(t)| + 10^{-8}\right]. \tag{162}$$

Gaussian OU dataset: temporally correlated trials. To introduce within-trial dependence without changing the time-local marginals, let $L_{\mathrm{G}}(t)L_{\mathrm{G}}(t)^\top = \Sigma_{\mathrm{G}}(t)$. For each condition c and trial r, standardized residuals follow the stationary OU recursion

$$z_{c,r,k} = \rho z_{c,r,k-1} + \sqrt{1-\rho^2}\,\eta_{c,r,k}, \qquad k = 2, \dots, T, \tag{163}$$

$$\rho = e^{-\Delta t/2} = e^{-0.5/2}, \qquad z_{c,r,1}, \eta_{c,r,k} \overset{\text{ind}}{\sim} \mathcal{N}(0, I_d). \tag{164}$$

The observations are

$$x_{c,r,k} = s_c\, 0.7\mu(t_k) + L_{\mathrm{G}}(t_k) z_{c,r,k}. \tag{165}$$

Each trial is an independent time sequence. Affine Flow uses a condition-balanced, trajectory-disjoint 80/20 training/validation split, so all time points from a trial remain in the same split. Bin + LW and GKR use all available trials.

Error is summarized by

$$\mathrm{MRAE} = \frac{1}{T}\sum_{k=1}^{T} \frac{\left|\widehat{D}_{\mathrm{J}}(t_k) - D_{\mathrm{J}}(t_k)\right|}{|D_{\mathrm{J}}(t_k)|}. \tag{166}$$

We compare Bin + LW, GKR, Affine Flow, and the Wishart process. The representative case has $d = 60$ and $R = 260$ trials per condition. The trial sweep fixes $d = 60$ and uses $R \in \{80, 215, 350, 485, 620\}$; the dimension sweep fixes $R = 260$ and uses $d \in \{20, 40, 60, 80, 100\}$.

Non-Gaussian two-scale OU dataset. For each condition c and trial r, a narrow or wide covariance branch $S_{c,r} \in \{-1, +1\}$ is drawn with equal probability and held fixed over the complete trajectory. Its time-local conditional distribution and base covariance are

$$X \mid c, t, S_{c,r} \sim \mathcal{N}(s_c\, 0.4\mu(t), [1 + S_{c,r}\rho_c(t)]\Sigma_{\mathrm{M}}(t)), \tag{167}$$

$$[\Sigma_{\mathrm{M}}(t)]_{j\ell} = 4\delta_{j\ell}\left[d\left(\frac{0.2}{\sqrt{2}}\right)^2 + \frac{d}{50}\,0.65|\mu_j(t)| + 10^{-8}\right]. \tag{168}$$

The two conditions have opposite time modulation,

$$\rho_0(t) = 0.5 + 0.4\sin[\omega(t+6) + 2.2], \qquad \rho_1(t) = 0.5 - 0.4\sin[\omega(t+6) + 2.2], \qquad \omega = \frac{2\pi}{12}. \tag{169}$$

Thus, with $L_{\mathrm{M}}(t)L_{\mathrm{M}}(t)^\top = \Sigma_{\mathrm{M}}(t)$, the complete trajectory is

$$x_{c,r,k} = s_c\, 0.4\mu(t_k) + \sqrt{1 + S_{c,r}\rho_c(t_k)}\, L_{\mathrm{M}}(t_k) z_{c,r,k}. \tag{170}$$

Equal branch probabilities preserve the base mean and covariance, while the opposite modulations make higher-order moments condition dependent. The exact full-mixture Jeffreys reference is Monte Carlo audited.

We compare GKR, Affine Flow, the Wishart process, and Unconstrained Flow. The representative case has $d = 10$ and $R = 260$ trials per condition. The trial sweep fixes $d = 10$ and uses $R \in \{100, 260, 420, 580, 740\}$; the dimension sweep fixes $R = 260$ and uses $d \in \{2, 10, 18, 26, 34\}$. Both sweeps use five repeats.

D.3 Real neural datasets

D.3.1 Mouse V1 static-grating dataset

The mouse V1 static-grating dataset contains six two-photon calcium-imaging sessions (GT1, GT2, GT3, TX38, TX39, and TX40) collected during full-field static-grating presentation (Stringer et al., 2021). Each trial contributes one population-response vector and one grating orientation, treated modulo π; each session contains about 4000 trials and roughly 20000 simultaneously recorded neurons. Within each session, we form a deterministic, orientation-stratified 64/16/20% training/validation/test split using seed 7. We fit a centered, unwhitened PCA using only the training trials, retain 200 components, and apply the fixed projection to validation and test trials. We chose 200 because it is close to the across-session median of 209 components required to reach a 70% cumulative explained-variance ratio.

We compare five conditional density estimators on the same splits: binning with Ledoit–Wolf covariance (Bin+LW), GKR, the Wishart process, affine flow, and unconstrained flow. Bin+LW uses 16 equal orientation bins of width $\pi/16$, GKR follows Section D.1.3, and the Wishart process follows Section D.1.4, using the adaptive 16-bin fitting protocol and a full-rank covariance factorization. Both flow models use eight periodic radial-basis features for orientation and a three-hidden-layer MLP of width 256; their fitting follows Section D.1.1.

We evaluate held-out log likelihood on the test split. For Bin+LW, GKR, the Wishart process, and affine flow, we use the fitted mean and covariance to compute Gaussian log likelihood; for unconstrained flow, we use the continuous change-of-variables identity in Eq. (111). This comparison measures conditional density fit rather than direct Fisher-information accuracy.

D.3.2 Allen Brain Observatory two-photon calcium-imaging dataset with drifting gratings

We use two-photon calcium-imaging responses from 11 Allen Brain Observatory visual-cortex recording sessions selected to contain more than 300 recorded neurons (de Vries et al., 2020). Mice viewed drifting gratings spanning eight motion directions ($0°, 45°, \ldots, 315°$) and five temporal frequencies ($1, 2, 4, 8$, and 15 Hz), giving 40 direction–frequency conditions. Each condition has 14–15 repeated trials, corresponding to approximately 560–600 nonblank trials per session. A trial contains 120 frames sampled at 30 Hz and spans four seconds: one second before stimulus onset, two seconds of drifting-grating presentation, and one second after stimulus offset, which also forms the first prestimulus second of the next trial.

For each session, we pool all nonblank observations (conditions $\times$ trials $\times$ time points) and randomly split into 60% training, 20% validation, and 20% test sets. Trial identity is not used in this split, so frames from the same trial can occur in multiple subsets. We fit a centered, unwhitened PCA to the raw training observations only, retain 95 components, and apply this fixed map to all observations. PCA-95 retains 60.1–89.6% of the training variance across sessions (median 70.2%).

The resulting 95-dimensional observations are conditioned on a categorical index for the 40 direction–frequency combinations and on continuous time. We compare Bin + LW, GKR, the Wishart process, affine flow, and unconstrained flow. Training responses are used for fitting, validation responses for checkpoint or hyperparameter selection, and test responses only for final evaluation.

Supplementary Figure S11 summarizes the responses, held-out likelihood, and time-resolved distance estimates across the 11 two-photon sessions.

D.3.3 Allen Visual Coding Neuropixels dataset

We analyze all 32 recording sessions from the Allen Visual Coding Neuropixels release (Siegle et al., 2021). These sessions share 40 nonblank drifting-grating conditions formed by eight motion directions and five temporal frequencies. For each session, we retain presentations whose $[-1, 3)$-second analysis window does not overlap an invalid interval. We retain units whose channel lies in VISp, VISl, VISrl, VISam, or VISpm and require presence ratio ≥ 0.95, inter-spike-interval violations ≤ 0.5, and amplitude cutoff ≤ 0.1. The resulting datasets contain 92–373 units and 421–600 presentations per session, with 8–15 presentations per condition.

We count spikes in 40 nonoverlapping 100-ms bins spanning $[-1, 3)$ seconds relative to stimulus onset and apply the elementwise square-root transform $k \mapsto \sqrt{k}$. Within each session, we fit a centered, unwhitened PCA to all concatenated condition–presentation–time observations before splitting and retain 70 components. These components explain 62.4–95.2% of the variance across sessions (median 72.9%). This is the space we are working on for the rest of analysis. We then randomly divide observation rows into 80% training and 20% test sets; the flow models further divide the outer training rows into 80% training and 20% validation rows.

The models condition on the 40-level direction–frequency index and continuous time. We compare condition-specific Bin + LW, GKR, affine flow, and unconstrained flow. Held-out likelihood is averaged within each condition and then across the 40 conditions, with recording session as the unit of aggregation. For Supplementary Figure S7, we average the 28 unique Jeffreys divergences among the eight directions separately at each temporal frequency within each recording session. Curves and shaded bands show the mean and SEM across the 32 sessions, respectively, with recording session as

the aggregation unit. The 4-Hz direction RDMs in Figure 7C are evaluated at the bin centers nearest -0.5, 1.0, and 2.5 seconds (realized centers -0.55, 0.95, and 2.45 seconds) and averaged across sessions.

D.4 Estimator and dataset for the geometry-template Jeffreys distance

Synthetic dataset. Let $s_\ell(U)$ parameterize the boundary of a centered square with side length ℓ, where $U \sim \mathrm{Unif}[0, 4]$. For condition c, we draw

$$X_c = R_{\theta_c} s_2(U) + \eta, \qquad \eta \sim \mathcal{N}(0, 0.2^2 I_2), \qquad (\theta_1, \theta_2) = (0, \pi/4). \tag{171}$$

Only X_c and its condition angle θ_c are included in the observed dataset; the perimeter coordinate U is latent. We generate 200 observations per condition and split each condition into 128 training, 32 validation, and 40 test observations (64/16/20%).

Flow-matching fitting. Both models start from the same side-length-one square boundary $s_1(U)$ convolved with isotropic Gaussian noise of initial scale 0.1; thus, the clean source geometry is identical across conditions. We compare two velocity families, each parameterized by a two-hidden-layer MLP of width 64. The similarity-constrained model outputs translation, angular velocity, isotropic scale, and center parameters, so its endpoint map can only translate, rotate, and uniformly rescale the square. The unconstrained model instead maps the response, flow time, and condition directly to a free two-dimensional velocity and may deform the template nonlinearly. We first train each model by Flow Matching along a cosine interpolation path using batch size 256, learning rate 10^{-3}, gradient clipping at norm 10, and validation early stopping with patience 1000, while holding the source noise fixed. We then fine-tune by negative log likelihood for 1200 epochs at learning rate 10^{-4}, allow a learnable source noise scale σ_c, and retain the checkpoint with the lowest validation NLL.

E SUPPLEMENTARY FIGURES

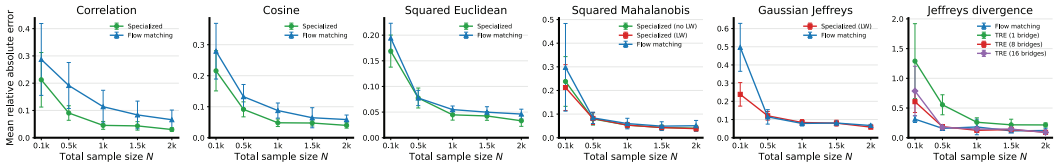

Figure S1: **Estimator accuracy across six representational distances on the five-condition categorical benchmark.** Mean relative absolute RDM error is shown against total sample size for correlation, cosine, squared Euclidean, squared Mahalanobis, Gaussian Jeffreys divergence, and full Jeffreys divergence. Specialized estimators are compared with flow matching; the full-Jeffreys panel also includes telescoping density-ratio estimation (TRE) with 1, 8, or 16 bridges (Rhodes et al., 2020). Markers and error bars show the mean and one sample standard deviation across five repetitions.

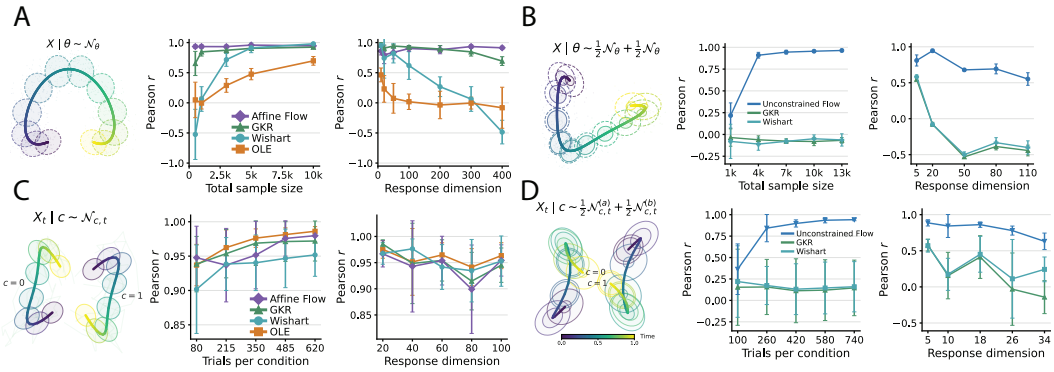

Figure S2: **Pearson correlation between the estimated and ground-truth curves.** (**A**) Independent Gaussian responses and linear Fisher information. (**B**) Independent two-scale Gaussian-mixture responses and full Fisher information. (**C**) Gaussian Ornstein–Uhlenbeck trials and time-resolved Gaussian Jeffreys divergence. (**D**) Two-scale Gaussian-mixture Ornstein–Uhlenbeck trials and time-resolved full Jeffreys divergence. Method colors follow Figures 4 and 6; points and error bars show the mean and one sample standard deviation across five repetitions. Figure relates to Figures 4 and 6

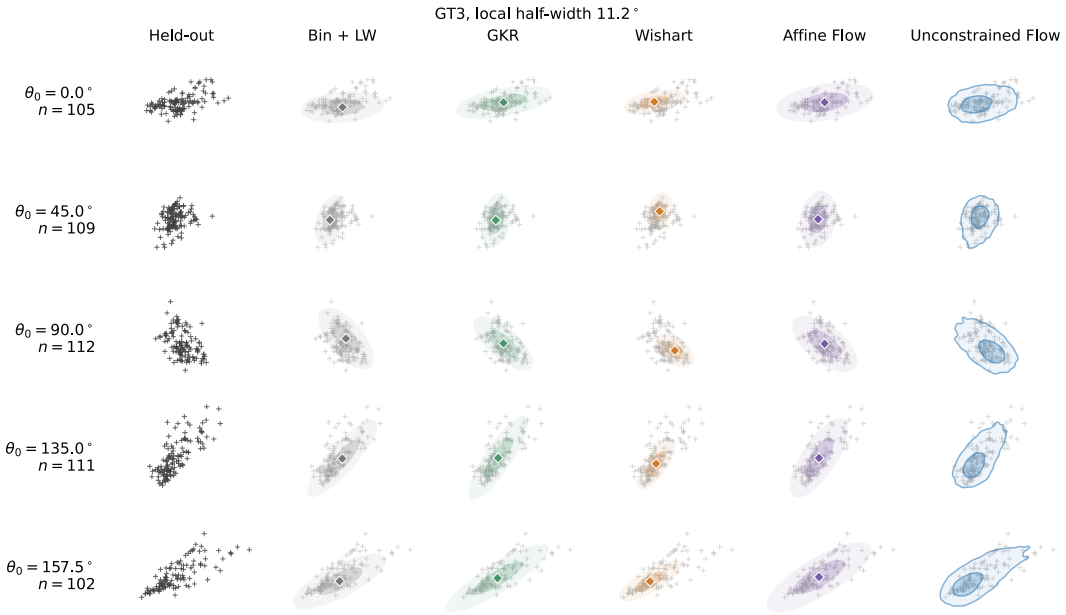

Figure S3: **Local conditional density fits across orientations in the representative Stringer session GT3.** Rows show held-out responses in the first two principal components within circular $\pm 11.25°$ windows centered at the displayed orientations. Columns compare held-out data, Bin + LW, GKR, the Wishart process, affine flow, and unconstrained flow. Crosses denote held-out observations; inner and outer ellipses or contours enclose 50% and 95% probability mass, respectively.

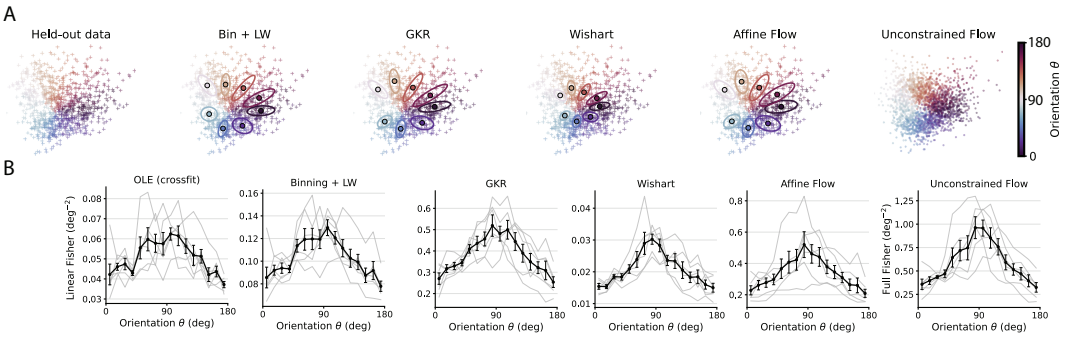

Figure S4: **Conditional density and Fisher-information estimates for the Stringer recordings.** (**A**) Held-out responses and fitted conditional densities in the first two principal components for an example session, colored by grating orientation. Gaussian estimators show fitted means and probability ellipses, whereas the unconstrained-flow panel shows generated samples. (**B**) Fisher-information estimates across the six sessions. GKR, Bin + LW, cross-fitted OLE, the Wishart process, and affine flow estimate linear Fisher information; unconstrained flow estimates full Fisher information. Thin gray curves show individual sessions, and colored curves and error bars show the pointwise mean and session-level standard error of the mean.

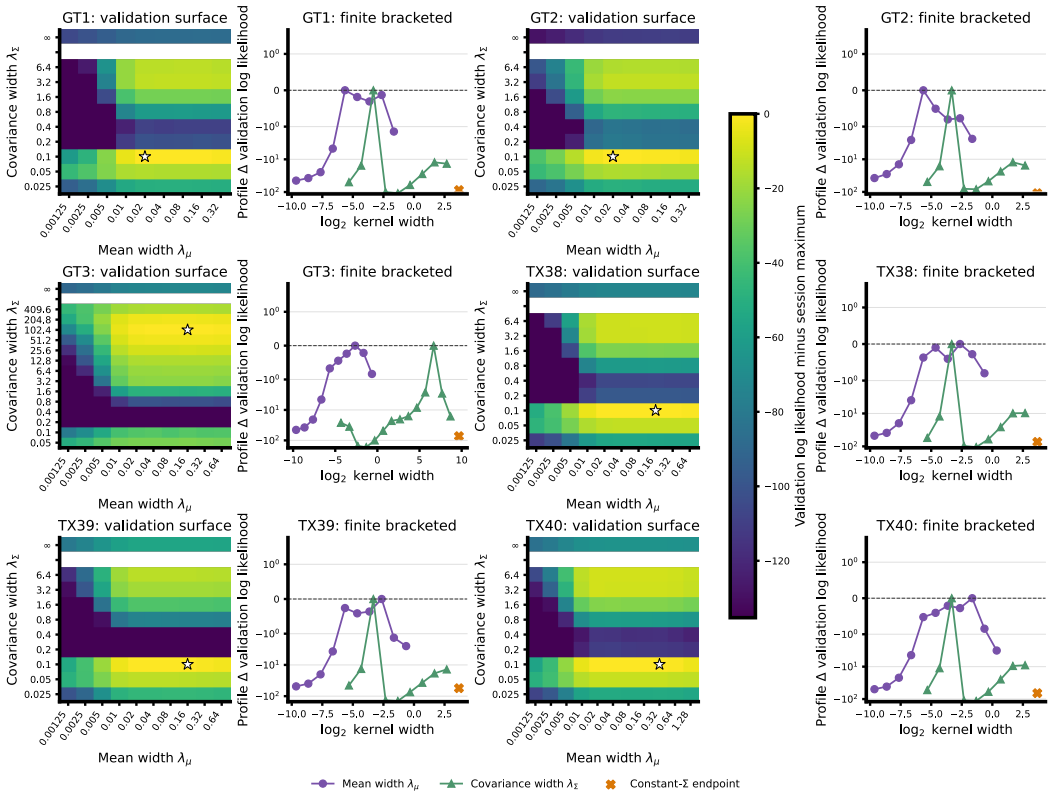

Figure S5: **Adaptive Wishart kernel-width searches for the Stringer recordings.** For each session, the validation surface shows Gaussian log likelihood relative to the session maximum over the mean width λ_μ and covariance width λ_Σ; the white star marks the selected pair. The adjacent profile panel shows relative validation likelihood across mean widths (purple circles) and finite covariance widths (green triangles), with the constant-covariance endpoint shown by the orange cross.

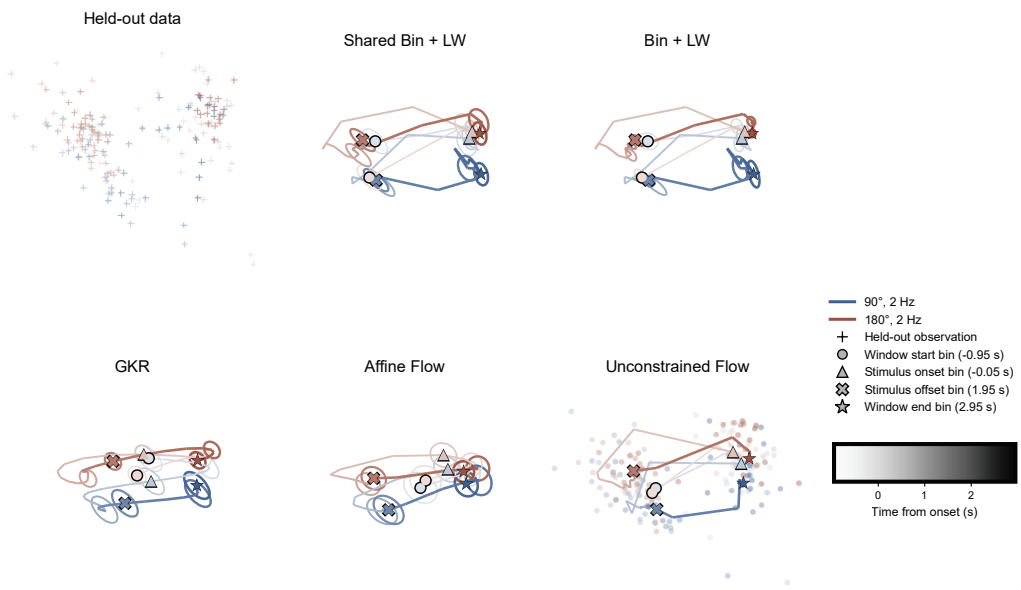

Figure S6: **Time-resolved conditional density trajectories for two drifting-grating conditions in the Allen Neuropixels recordings.** Held-out responses and fitted distributions are shown in the first two principal components for the $(90°, 2\,\mathrm{Hz})$ and $(180°, 2\,\mathrm{Hz})$ conditions. Columns show held-out data, shared Bin + LW, Bin + LW, GKR, affine flow, and unconstrained flow. Marker shape identifies representative time bins from the beginning to the end of the analysis window, and color lightness indicates time from stimulus onset.

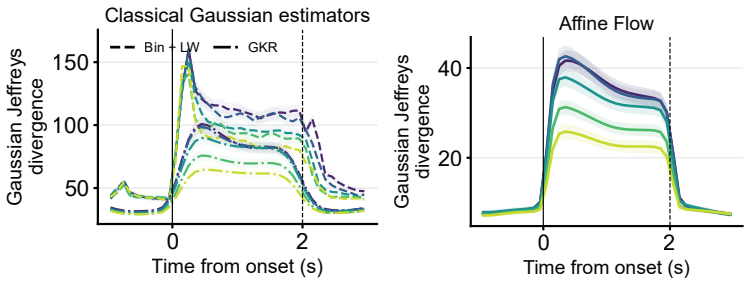

Figure S7: **Frequency-specific time-resolved Gaussian Jeffreys divergence from classical estimators and affine flow across 32 Allen Neuropixels sessions.** At each temporal frequency, Gaussian Jeffreys divergence is averaged over the 28 unique pairs among the eight directions within each session. Colors indicate the five temporal frequencies (1, 2, 4, 8, and 15 Hz). The left panel shows Bin + LW (dashed curves) and GKR (dash-dotted curves); the right panel shows affine flow (solid curves). Curves show across-session means, shaded regions show the SEM, and solid and dashed vertical lines mark stimulus onset and offset. Bin + LW uses a 0.2-s time window.

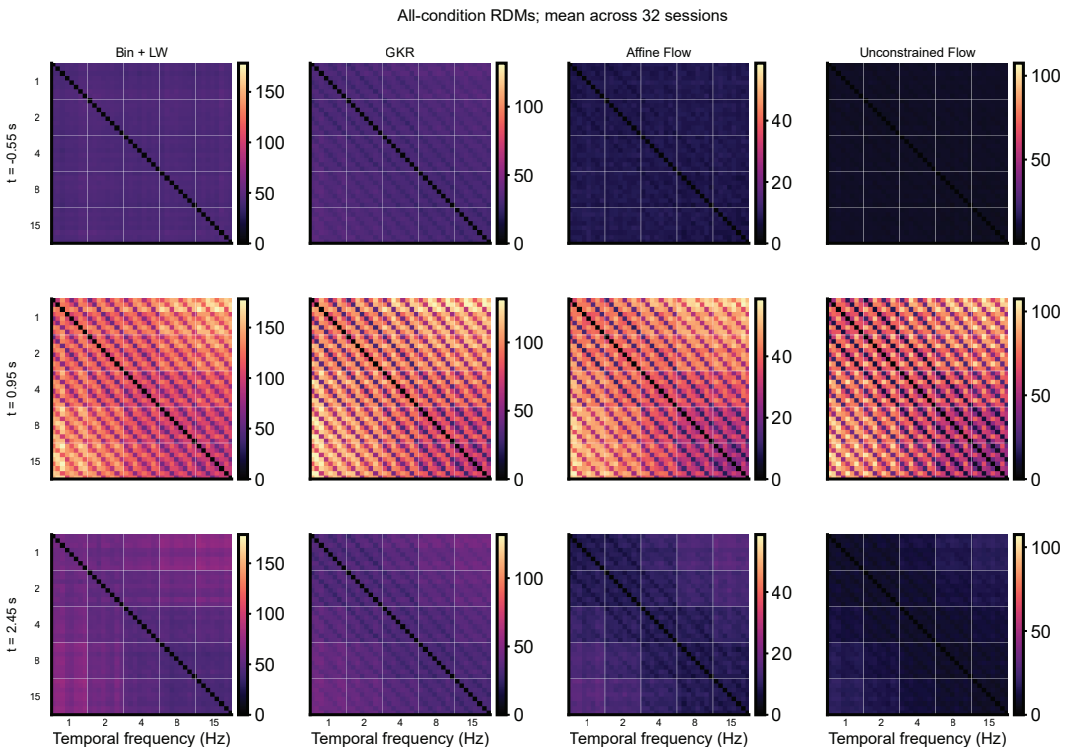

Figure S8: **Time-resolved all-condition RDMs across 32 Allen Neuropixels sessions.** Each matrix is the elementwise mean across sessions for all 40 direction–frequency conditions at -0.55, 0.95, or 2.45 seconds relative to stimulus onset. Columns show Gaussian Jeffreys divergence from Bin + LW, GKR, and affine flow, followed by full Jeffreys divergence from unconstrained flow. Conditions are grouped by temporal frequency, with eight motion directions within each block.

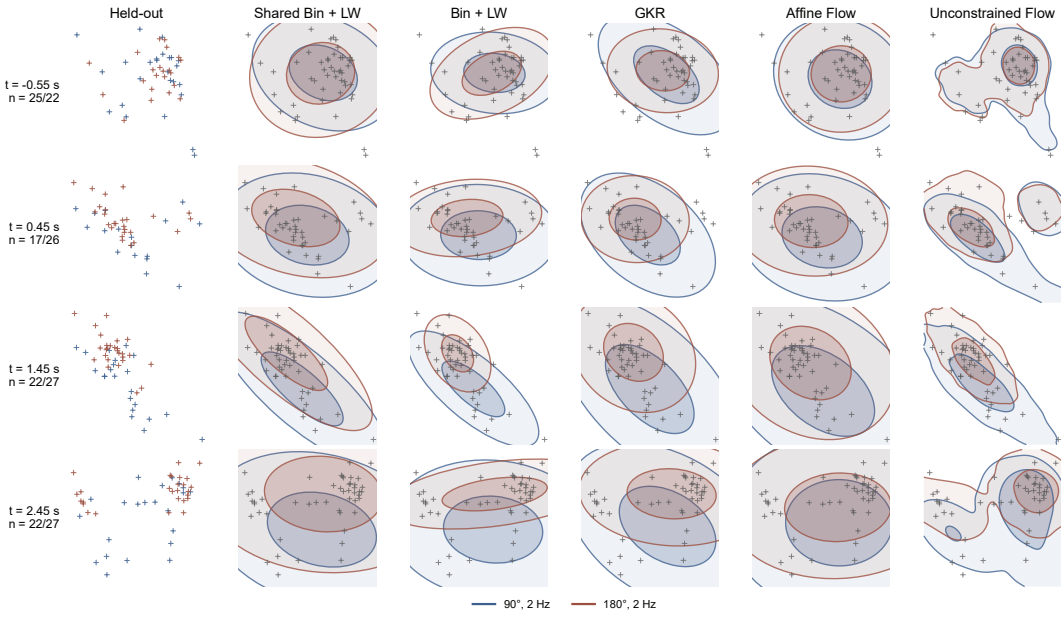

Figure S9: **Conditional density fits at four example times in an example Allen Neuropixels session.** Held-out responses and fitted conditional densities are shown in the first two principal components for the $(90°, 2\,\mathrm{Hz})$ and $(180°, 2\,\mathrm{Hz})$ conditions. Rows correspond to windows centered at -0.55, 0.45, 1.45, and 2.45 seconds relative to stimulus onset; the displayed counts give the held-out observations for the two conditions. Columns compare held-out data, shared Bin + LW, Bin + LW, GKR, affine flow, and unconstrained flow.

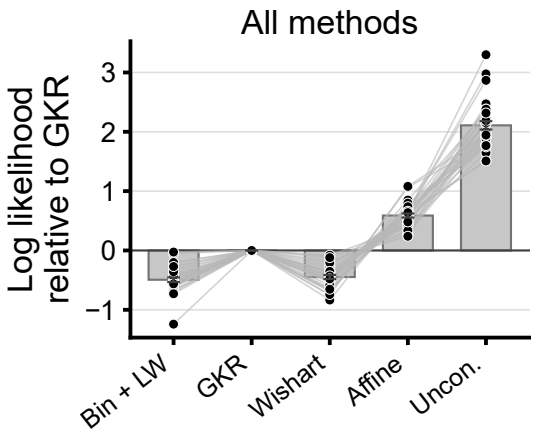

Figure S10: **Held-out likelihood across five Allen Neuropixels sessions after projection onto 10 principal components.** Held-out log likelihood relative to GKR is shown for Bin + LW, GKR, the Wishart process, affine flow, and unconstrained flow. Gray lines connect estimates from the same session, black dots show individual sessions, and bars and error bars show the across-session mean and standard error of the mean.

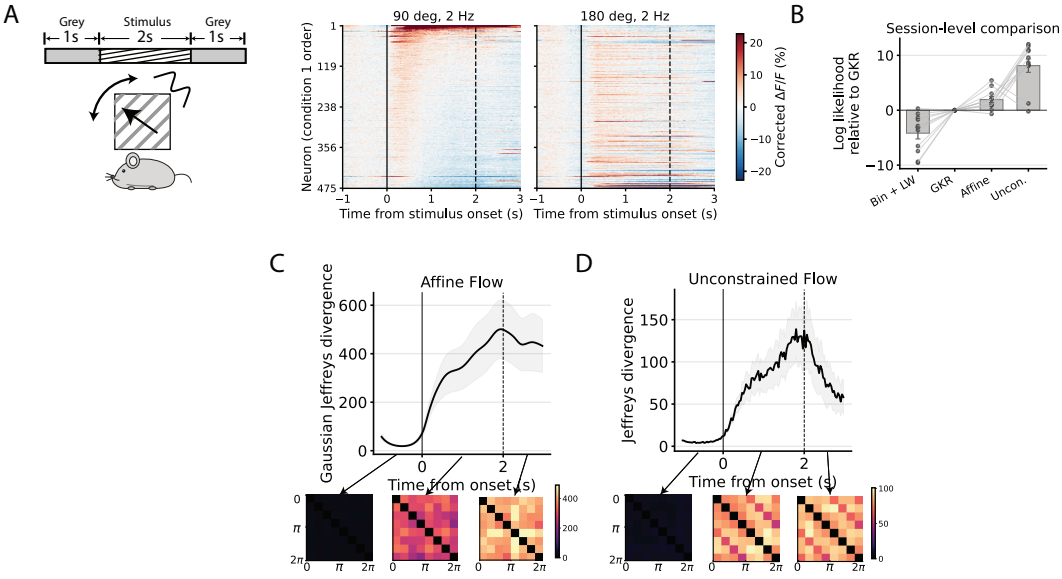

Figure S11: **Responses, held-out likelihood, and time-resolved representational geometry in the Allen two-photon recordings.** (**A**) Four-second trial structure and baseline-corrected response heat maps for two drifting-grating conditions in a representative session. (**B**) Held-out log likelihood relative to GKR across 11 imaging sessions; gray lines connect estimates from the same session, and bars and error bars show the across-session mean and standard error of the mean. (**C**) Affine-flow Gaussian Jeffreys divergence averaged over condition pairs and 4-Hz direction RDMs at representative pre-stimulus, stimulus, and post-stimulus times. (**D**) Corresponding full Jeffreys divergence and RDMs from unconstrained flow. Curves and shaded regions summarize mean and SEM of the 11 sessions.

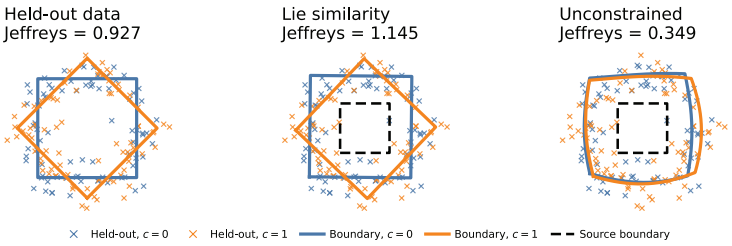

Figure S12: **Similarity-constrained flow preserves a square template when estimating representational dissimilarity.** Left: held-out observations from two noisy-square conditions, with solid curves showing their noiseless target boundaries. Middle and right: fits obtained using similarity-constrained and unconstrained flows, respectively, with the same side-length-one square source. In the fitted panels, the dashed black curve shows the original source boundary, and the solid colored curves show the condition-specific transported boundaries. Crosses denote held-out observations, and each panel title reports the Jeffreys divergence between the two condition distributions. The similarity-constrained fit preserves the square geometry, whereas the unconstrained fit distorts it.

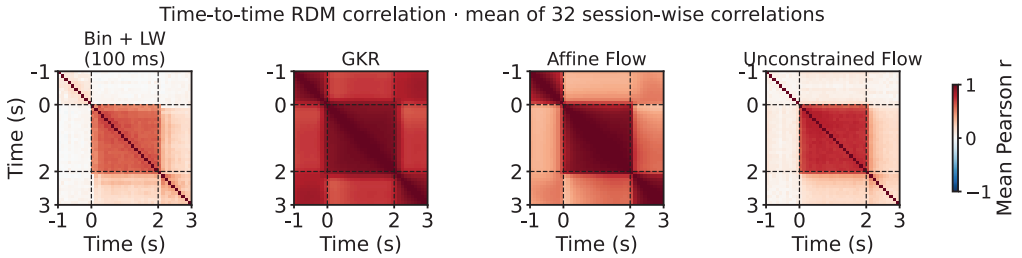

Figure S13: **Time-to-time RDM correlations across estimators in the Allen Neuropixels recordings.** Panels compare Bin + LW, GKR, affine flow, and unconstrained flow. Each matrix shows Pearson correlations between RDMs at pairs of time points, computed within each recording session and then averaged across 32 sessions. Dashed lines mark stimulus onset at 0 seconds and offset at 2 seconds.