# INTRINSIC GEOMETRY MATCHING: LEARNING OBSERVATION-EQUIVARIANT PROCESS GEOMETRY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Real data are often arbitrary, indirect observations of an unknown underlying process, and those observations can distort its geometry. Existing geometric methods equip data with the induced geometry of the ambient space, but this entangles the true geometry of the process with the distortion of the observation. To address this, we introduce **Intrinsic Geometry Matching**: a scalable approach to learning the intrinsic geometry of the underlying process. Specifically, we learn the *carré du champ* operator of the process, allowing us to recover its diffusion geometry. Our method uses the fact that data often come with natural perturbations, like steps in a time series, to recover the intrinsic carré du champ while transforming naturally with changes in observation: it is **observation-equivariant**. Furthermore, since this equivariance relates the geometry of different spaces, we can train a model using additional sources of data, even when that data lies in another space. We evaluate IGM on molecular dynamics, image, and video data and find better performance and scaling than existing neural and graph-based geometric tools.

## 1 INTRODUCTION

Real data are often arbitrary, indirect observations of an unknown underlying process. In biology, cells, tissues, genomes, and proteins evolve and adapt, but can only be observed through noisy and often destructive experiments. Climate and weather systems have complex, chaotic structure, which is only partially captured by a sparse network of sensors and satellites. These processes are governed by physical laws and constraints, so the data we collect have *geometry*. This geometry can tell us about the structure of the underlying system, but the geometry we observe is distorted by the nature of the observation. For example, the distance between two satellite images as vectors is not the same as the physical 'distance' between the weather systems they observe. As the frontier of scientific AI moves from representing states to representing processes, the corresponding geometric challenge is to disentangle the true **process geometry** from the arbitrary **observation geometry**.

Scientific modelling is fundamentally concerned with systems under change: trajectories, perturbations, constraints, interventions, and controls. The mathematical language of change is **calculus**, which works with notions of derivative. On nonlinear manifolds, this is generalised by **Riemannian geometry**, whose basic primitive is the Riemannian metric. This picture provides a useful geometric intuition for machine learning as the *Manifold Hypothesis*, which posits that high-dimensional data actually lie on a low-dimensional manifold (Tenenbaum et al., 2000; Roweis and Saul, 2000; Bengio et al., 2012). In reality, it is mostly just a metaphor: many complex scientific systems exhibit varying dimensionality, branches, singularities and discrete–continuous structure, so a smooth manifold of fixed dimension is too rigid an operational model.

To properly model this heterogeneous structure, we need a geometric theory that is both more flexible and computable from data. These requirements point to **diffusion geometry** (Jones, 2024; Bakry et al., 2013), which extends Riemannian constructions to irregular manifold-like structures via the calculus of Markov diffusion operators. Since it is defined in terms of a process, it provides a natural model for the *process geometry* underlying observed data. Its metric-like analogue is the carré du champ (CDC) operator (Bakry and Émery, 1985; Kunita, 1969), which measures the local directions of variability within the space, and can be used to recover its calculus and geometry. Crucially, the CDC transforms naturally under observation: when an underlying, latent process is observed by a map into $\mathbb{R}^D$, the CDC pushes forward to $\mathbb{R}^D$ by conjugation with the derivative of the map. We



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

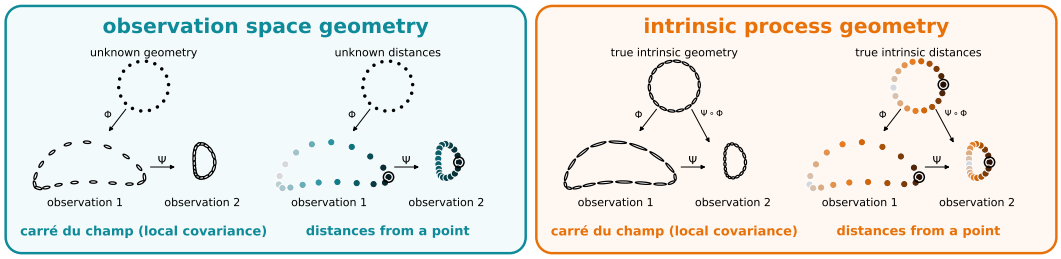



Figure 1: **Observation equivariance.** Standard geometric tools give data the **geometry of the observation space (left)**, so depend on the observation map. This affects the carré du champ, which measures local covariance and describes the data geometry, like the distances. These quantities cannot be reconciled across different observations. We learn the **intrinsic geometry of the underlying process (right)**. The carré du champ measures intrinsic covariance and is observation-equivariant: it transforms naturally when we change observation. Distances are intrinsic and consistent.

call this relationship **observation equivariance**: the geometry is still intrinsic, and the observation map can be exchanged for another in a predictable way (fig. 1, right). If we could find a way to learn this natural CDC from data, we would have successfully disentangled the process geometry from the observation geometry via a clean algebraic factorisation.

Standard geometric methods treat the data as an unordered point cloud with no additional structure, and equip it with the geometry of the ambient space it is observed in (Lebanon, 2012; Hauberg et al., 2012; Berry and Giannakis, 2020; Jones and Lanners, 2026). On a manifold, this means using the induced ambient Riemannian metric, so geometric measurements, like geodesic distances, are observation-dependent (fig. 1, left). While this problem can be partially addressed with graph-based methods when the Jacobian of the observation map is known or can be estimated from dense local samples (Singer and Coifman, 2008; Singer et al., 2009; Talmon and Coifman, 2013), this information is rare in practice. Graph methods also suffer from drawbacks like poor scalability and generalisation, and the curse of dimensionality, which can be overcome with deep learning.

Our approach overcomes these challenges by exploiting the fact that many real data sets have additional semantic structure in the form of natural *perturbations* or *augmentations*, which define their diffusion geometry. For example, the steps in a time series represent a *genuine sample from the underlying process*. We show that the true CDC, and hence the true geometry, can be inferred from these perturbations, and provide a scalable algorithm for training a neural network to learn it.

**Main contributions:**

- C1. When data come from an unknown, latent stochastic process, we describe the intrinsic geometry in terms of the carré du champ operator (CDC). This CDC is *intrinsic to the process* and *observation-equivariant*.
- C2. We present **Intrinsic Geometry Matching (IGM)**: a scalable loss for learning the intrinsic CDC from data perturbations, and show that it is the unique observation-invariant loss.
- C3. We exploit observation equivariance to train IGM from data in different spaces.
- C4. We evaluate IGM on a range of high-dimensional datasets, demonstrating better performance, scaling properties, and training stability than existing methods.

## 2 Background

### 2.1 Theories of differential geometry

Let $\mathcal{M}$ be a smooth $d$-dimensional manifold. A **Riemannian metric** $g$ assigns, at each point $x \in \mathcal{M}$, an inner product $g_x$ on the tangent space $T_x\mathcal{M}$. Its inverse $g_x^{-1}$ is an inner product on the cotangent space $T_x^*\mathcal{M}$ called the **cometric**. These allow calculus objects to be defined on manifolds with curvature. If $\Psi : \mathcal{M} \to \mathcal{N}$ is a smooth map with derivative $D_x\Psi$ at $x \in \mathcal{M}$, we can use $D_x\Psi$ to transport the metric and cometric back and forth between $\mathcal{M}$ and $\mathcal{N}$. The metric $g_{\mathcal{N}}$ *pulls back* to $(D_x\Psi)^\top g_{\mathcal{N}}(\Psi(x))\,(D_x\Psi)$, and the cometric $g_{\mathcal{M}}^{-1}$ *pushes forward* to $(D_x\Psi)g^{-1}(x)\,(D_x\Psi)^\top$.

**Diffusion geometry** extends Riemannian geometry to non-manifolds using the Bakry–Émery theory of Markov diffusion operators (Jones, 2024; Bakry and Émery, 1985). On a manifold $\mathcal{M}$, the geom-

etry and calculus are captured by the metric, which satisfies $2g(\nabla f, \nabla h) = \Delta(fh) - f\Delta h - h\Delta f$, where $\Delta$ is the Laplace-Beltrami operator. Since $\Delta$ is the generator of Brownian motion on $\mathcal{M}$, this identity relates the geometry to a diffusion process. Diffusion geometry exploits this relationship to define geometry and calculus in more general settings, replacing the Brownian motion with a general Markov diffusion. Let $(X_\tau)_{t\geq 0}$ be a Markov process on a measure space $\mathcal{X}$. If $f : \mathcal{X} \to \mathbb{R}$, its *transition semigroup* is the operator $f \mapsto P_t f(x) := \mathbb{E}[f(X_\tau) \mid X_0 = x]$, and its *infinitesimal generator* is $\mathcal{L}f(x) := \lim_{t\downarrow 0} t^{-1}(P_t f(x) - f(x))$. We define the ***carré du champ* (CDC) operator**

$$\Gamma_{\mathcal{L}}(f,h)(x) := \tfrac{1}{2}\left(\mathcal{L}(fh)(x) - f(x)\mathcal{L}h(x) - h(x)\mathcal{L}f(x)\right) \tag{1}$$

$$= \lim_{\tau\to 0} \tfrac{1}{2\tau}\mathbb{E}[(f(X_\tau) - f(X_0))(h(X_\tau) - h(X_0)) \mid X_0 = x], \tag{2}$$

so that, when $\mathcal{L} = \Delta$ on a manifold, $\Gamma_\Delta(f,h) = g(\nabla f, \nabla h)$ (Kunita, 1969; Bakry and Émery, 1985). In local coordinates $x_i$, the matrix $[\Gamma_{\mathcal{L}}(x_i, x_j)]_{ij} = [g(\mathrm{d}x_i, \mathrm{d}x_j)]_{ij}$ is exactly the *cometric*. For diffusion processes it satisfies the *diffusion property* (chain rule): if $\Phi$ is differentiable, then $\Gamma_{\mathcal{L}}(\Phi(f_1,\dots,f_k), h) = \sum_{i=1}^k \partial_i \Phi(f_1,\dots,f_k)\,\Gamma_{\mathcal{L}}(f_i, h)$ (Bakry et al., 2013).

### 2.2 Forward and backward Markov processes

The methods we discuss here all use a notion of local kernel or perturbation, which can be unified in the language of *Markov processes*. Let $\mu$ be a probability measure on a state space $\mathcal{X}$, and let $P^+(x, \mathrm{d}y)$ be a **Markov kernel**: a probability measure for perturbations $y$ of a sample $x$. We call $P^+$ the **forward process**. If $X \sim \mu$ and $Y \sim P^+(X, \cdot)$, this defines a joint distribution $\pi(\mathrm{d}x, \mathrm{d}y)$ and marginal $\nu(\mathrm{d}y)$ for $Y$. Together, these define the **backward process** $P^-(y, \mathrm{d}x)$ by $\mu(\mathrm{d}x)\,P^+(x, \mathrm{d}y) = \pi(\mathrm{d}x, \mathrm{d}y) = \nu(\mathrm{d}y)\,P^-(y, \mathrm{d}x)$, which represents the conditional distribution of $X$ given $Y = y$, so $P^-$ is the $\mu$-dependent *posterior* given the *prior* $P^+$.

### 2.3 Computing diffusion geometry

Classical geometric methods use **graphs** and **kernels** to measure the affinity between points. For example, if $k_\tau$ is a symmetric, positive kernel with bandwidth $\tau$, then the weighted graph Laplacian of *Laplacian eigenmaps* (Belkin and Niyogi, 2003) and *diffusion maps* (Coifman and Lafon, 2006) is given by $\Delta_\tau(f)(y) = \tau^{-1}\sum_{i=1}^N p_\tau(x_i, y)(f(y) - f(x_i))$ where $p_\tau(x,y) = k_\tau(x,y)/\sum_{j=1}^N k_\tau(x_j, y)$. On a manifold with measure $\mu = \rho\mathrm{dvol}$, $\Delta_\tau(f)$ converges to $\Delta(f) - g(\nabla \log\rho, \nabla f)$ as $\tau \to 0$ (Coifman and Lafon, 2006). Laplacian estimates can be used to estimate the CDC: if $\Delta_\tau(f)$ is plugged into equation 1, we obtain the covariance-like formula

$$\Gamma_\tau(f,h)(y) = \frac{1}{2\tau}\sum_{i=1}^N p_\tau(x_i, y)(f(y) - f(x_i))(h(y) - h(x_i)), \tag{3}$$

with convergence $\Gamma_\tau(f,h) \to \Gamma(f,h)$ as $\tau \to 0$. This has been used for kernel-based computation of diffusion geometry (Jones, 2024; 2026; Berry and Giannakis, 2020). For a more accurate and robust estimator, we can estimate the CDC as the true covariance $\Gamma_\tau(f,h)(y) = (2\tau)^{-1}\mathrm{Cov}[f(X), h(X) : X \sim p_\tau(\cdot, y)]$ (Bamberger et al., 2026b; Jones and Lanners, 2026).

However, these kernel-based approaches have all the drawbacks of kernel machine learning: poor scalability, the curse of dimensionality, and poor generalisation. This has motivated a range of new geometric methods based on **deep learning**. Neural networks can be trained to estimate kernel-based expressions by recognising that kernel functions, like the Gaussian, are *likelihood functions*, whose closed-form expressions can be replaced by Monte Carlo averages over samples. By finding an appropriate sample-wise loss, we can estimate the same objects without ever evaluating the kernel.

This idea is central to diffusion models, where *score matching* (Song et al., 2020; Song and Ermon, 2019) is a kernel-free estimate of *Laplacian eigenmaps*. In this special case of Gaussian noise, we can recover the geometry for free, since the covariance-based CDC formula above is just the Jacobian of the score matching denoiser (Manor and Michaeli, 2024; Kharitenko et al., 2026), and is also related to the second-order score (Meng et al., 2021). The first method to learn the CDC in full generality was *Riemannian metric matching (RMM)* (Bamberger et al., 2026a), which can target either of the CDC estimates above, and exploits a low-rank parametrisation to allow scaling to high-dimensional spaces.

## 3 Intrinsic Geometry of Observed Processes

Geometric methods involve a kernel or prior, and make estimates involving the posterior. We ask: if the prior is more naturally adapted to the data, is the learned geometry more meaningful? In the extreme: if the prior is genuinely *intrinsic*, like sampling steps in a time series, can we learn the true intrinsic geometry? We now define what 'intrinsic geometry' means for data.

We will use the language of diffusion geometry (Jones, 2024), and model the underlying process geometry as a diffusion Markov triple $(\mathcal{X}, \mu, \Gamma)$, where $\mathcal{X}$ is a measure space with measure $\mu$, and $\Gamma$ is the carré du champ (CDC) operator (Bakry et al., 2013). We make observations in $\mathbb{R}^D$ via some map $\Phi : \mathcal{X} \to \mathbb{R}^D$, so the data we see are samples $X = \Phi(\alpha)$ for $\alpha \sim \mu$, and we hope to recover the underlying process geometry. However, the observation map $\Phi$ is typically arbitrary: it is not an *isometric embedding* into $\mathbb{R}^D$, and can distort the structure.

**Recovering the intrinsic CDC in observation space.** Most geometric methods treat the observations as the *only information available*, and equip it with the **geometry of the ambient space** $\mathbb{R}^D$. For manifold data, this means giving the observed manifold the *induced* Riemannian metric (the dot product in $\mathbb{R}^D$). This geometry depends on $\Phi$, and is only correct when $\Phi$ is isometric. Alternatively, there is a **natural geometry** available (fig. 1). In the manifold case, this is given by the pushforward cometric $(D_x\Phi)g_x^{-1}(D_x\Phi)^\top$. Given a Markov triple, we can define the *pushforward CDC* on $\Phi_\sharp\mu$. If $\Phi$ is injective and if $f, h : \mathbb{R}^D \to \mathbb{R}$ and $x = \Phi(\alpha)$, then

$$\Phi_\sharp\Gamma(f, h)(x) := \Gamma(\Phi^* f, \Phi^* h)(\alpha), \tag{4}$$

where $\Phi^* f = f \circ \Phi$ denotes the pullback. If $\Phi$ is not injective, so there are multiple $x$ for which $y = \Phi(x)$, we can extend the definition to a *posterior average* over $x$, i.e.

$$\Phi_\sharp\Gamma(f, h)(x) := \mathbb{E}\big[\Gamma(\Phi^* f, \Phi^* h)(\alpha) \mid \Phi(\alpha) = x\big] \tag{5}$$

(Bakry et al., 2013, Sec. 1.15.4). Even though we remain in the observation space $\mathbb{R}^D$, with no knowledge of $\mathcal{X}$ or $\Phi$, we can use $\Phi_\sharp\Gamma$ to still work with the true geometry. For example, if geodesics (shortest paths) are computed in $\mathbb{R}^D$ using $\Phi_\sharp\Gamma$, they will be the pushforward of the true geodesics on $\mathcal{X}$. This definition is compositional: if $\Psi : \mathbb{R}^D \to \mathbb{R}^{D'}$ is differentiable, then $(\Psi\circ\Phi)_\sharp\Gamma = \Psi_\sharp\big[\Phi_\sharp\Gamma\big]$.

**The CDC of coordinate functions is sufficient and observation-equivariant.** At a point $y \in \mathbb{R}^D$, let $G_\Phi(y)$ be the $D \times D$ matrix of the CDC of the coordinate functions $x_i$, so

$$G_\Phi(x) = [\Phi_\sharp\Gamma(x_i, x_j)(x)]_{ij} = \mathbb{E}\big[\Gamma(\Phi_i, \Phi_j)(\alpha) \mid \Phi(\alpha) = x\big]_{ij}. \tag{6}$$

Notice that $G_\Phi$ is *sufficient to describe the whole CDC*, and hence all the geometry: if $f, h : \mathbb{R}^D \to \mathbb{R}$, then $\Phi_\sharp\Gamma(f, h)(x) = (D_x f)G_\Phi(x)(D_x h)^\top$. If the observation map $\Phi$ is replaced by the composition $\Psi \circ \Phi$ with an injection $\Psi$, and $y = \Psi(x)$ then the chain rule for $\Gamma$ gives

$$G_{\Psi\circ\Phi}(y) = (D_x\Psi)G_\Phi(x)(D_x\Psi)^\top, \tag{7}$$

a property we call **observation equivariance**: the CDC transforms naturally under changes in observation. This is desirable because it means that the geometry is measuring *only something intrinsic to the system*, independent of how it was observed. This defines the following organising principle.

> **Principle 1: intrinsic geometry is observation-equivariant.** Observations are usually non-canonical, so the learned geometry should be *observation-equivariant*. Equivalently, the CDC should be the pushforward of the *intrinsic* CDC to the observation space.

**Approximating $G_\Phi$ with the Markov process.** This ideal CDC $G_\Phi$ is defined as the infinitesimal limit of a Markov process as time $\tau \to 0$ (equation 2), and our method will approximate this limit at small $\tau > 0$. Let $(\alpha_\tau)_{\tau\geq 0}$ be a Markov process on $\mathcal{X}$ with CDC $\Gamma$, so we observe the **pushforward process** $X_\tau = \Phi(\alpha_\tau)$ with CDC $\Phi_\sharp\Gamma$. We define a covariance-style estimator

$$G_\tau(x) = \tfrac{1}{2\tau}\mathbb{E}\big[(X_\tau^i - X_0^i)(X_\tau^j - X_0^j) \mid X_0 = x\big]_{ij}, \tag{8}$$

so $\lim_{\tau\downarrow 0} G_\tau = G_\Phi$. We can also estimate $G_\Phi$ as the standard mean-centred covariance

$$G_{\tau,\mathrm{Cov}}(x) = \tfrac{1}{2\tau}\mathrm{Cov}\big[X_\tau^i, X_\tau^j \mid X_0 = x\big]_{ij} = \tfrac{1}{2\tau}\mathbb{E}\big[(X_\tau^i - \mathbb{E}[X_\tau]^i)(X_\tau^j - \mathbb{E}[X_\tau]^j) \mid X_0 = x\big]_{ij}. \tag{9}$$

If the underlying diffusion process has a *drift* $b$, then the mean $\mathbb{E}[X_\tau | X_0 = x]$ learns $X_0 + \tau b$, and centring removes the drift bias from the estimate and reduces the variance of the CBMM loss. By approximating either $G_\tau$ or $G_{\tau,\mathrm{Cov}}$, we can estimate $G_\Phi$ and recover the full CDC $\Phi_\sharp\Gamma$.

# 4 Intrinsic Geometry Matching

## 4.1 Conditional training with Bregman divergences

We now describe a procedure for training a neural network $G_\tau^\theta$ to learn $G_\tau$ or $G_{\tau,\mathrm{Cov}}$. Suppose we have samples of data $X \sim \mu$ and perturbations $Y \sim P^\tau(X, \mathrm{d}y)$. To train $G_\tau^\theta$, we need to pick a matrix loss function, and a natural choice is a *Bregman divergence*.

**Definition 4.1.** *Let $\mathcal{A} \subseteq \mathbb{S}^D$ be convex and $\omega : \mathcal{A} \to \mathbb{R}$ be strictly convex and differentiable. The **Bregman divergence** on $\mathcal{A}$ induced by $\omega$ is $\mathcal{D}_\omega(A, B) = \omega(A) - \omega(B) - \langle \nabla\omega(B), A - B\rangle$.*

For example, the Frobenius loss $\|A - B\|^2$ is the Bregman divergence induced by $\omega(X) = \|X\|^2$. Ideally, we would pick $\mathcal{D}(\cdot,\cdot)$, and regress $G_\tau^\theta$ over $\tau \sim \mathcal{U}[0, \tau_{\max}]$ and $X \sim \mu$ with

$$\mathcal{L}_{\mathrm{BMM}}(\theta) = \mathbb{E}_{\tau,X}\big[\mathcal{D}(G_\tau(X), G_\tau^\theta(X))\big],$$

which we call the **Bregman metric matching (BMM)** loss. Evaluating this loss is impossible as we do not know $G_\tau$ in advance, so we introduce the tractable alternative using $Y \sim P^\tau(X, \mathrm{d}y)$

$$\mathcal{L}_{\mathrm{CBMM}}(\theta) = \mathbb{E}_{\tau,X,Y|X}\big[\mathcal{D}(\tfrac{1}{2\tau}(Y - X)(Y - X)^\top, G_\tau^\theta(X))\big],$$

which we call the **conditional Bregman metric matching (CBMM)** loss. The CBMM loss can be cheaply evaluated given only samples $X$ and $Y$, and is trivially parallelisable over those samples for distributed training. Crucially, the simple and tractable $\mathcal{L}_{\mathrm{CBMM}}$ is equal to $\mathcal{L}_{\mathrm{BMM}}$ up to a constant independent of $\theta$, so $\nabla_\theta\mathcal{L}_{\mathrm{CBMM}} = \nabla_\theta\mathcal{L}_{\mathrm{BMM}}$ and training with either loss is equivalent.

**Proposition 4.2** (Conditional Bregman metric matching)**.** *We have, for sufficiently regular $\mathcal{D}$,*

$$\mathcal{L}_{\mathrm{CBMM}}(\theta) = \mathcal{L}_{\mathrm{BMM}}(\theta) + \mathbb{E}\big[\mathcal{D}(\tfrac{1}{2\tau}(Y - X)(Y - X)^\top, G_\tau(X))\big] \tag{10}$$

*if and only if $\mathcal{D}$ is a Bregman divergence (Banerjee et al., 2005, follows from Theorems 1 & 4).*

This shows that we can learn $G_\tau$ using *only samples from $P^\tau$*, without prior knowledge of a kernel function, so can apply our method to 'black box' data-driven perturbations. When $P^\tau$ represents intrinsic transitions of an unknown process, we can still recover the intrinsic geometry.

> **Principle 2: learn the intrinsic geometry from data-driven perturbations.** Data-driven perturbations (like steps in a time series) represent *oracle access to the intrinsic process*, and can be used to learn the intrinsic geometry despite the arbitrary observation.

**Mean-centred covariance.** As stated, CBMM learns the uncentred covariance $G_\tau$. To instead learn the mean-centred covariance $G_{\tau,\mathrm{Cov}}$, we need a model $m_\tau^\phi(x)$ for the mean $\mathbb{E}[X_\tau \mid X_0 = x]$, which can be trained with $\mathcal{L}(\phi) = \mathbb{E}_{\tau,X,Y|X}\|Y - m_\tau^\phi(X)\|^2$. We then define the conditional BMM loss

$$\mathcal{L}^{\mathrm{Cov}}_{\mathrm{CBMM}}(\theta) = \mathbb{E}_{\tau,X,Y|X}\big[\mathcal{D}(\tfrac{1}{2\tau}(Y - m_\tau^\phi(X))(Y - m_\tau^\phi(X))^\top, G_\tau^\theta(X))\big],$$

which, by the same argument as above, is equivalent to $\mathbb{E}_{\tau,X}[\mathcal{D}(G_{\tau,\mathrm{Cov}}(X), G_\tau^\theta(X))]$.

**Learning the backward process CDC.** We have defined the BMM and CBMM losses as estimators for the forward process CDC $G_\tau$, which is natural when $P^\tau$ gives samples from an underlying Markov process. However, most methods for computational geometry and generative modelling use an uninformative forward kernel $P^\tau$ (e.g. Gaussian), and the posterior $P^{-\tau}$ models the structure. We can learn the backward CDC $G_{-\tau}$ by simply replacing $X \sim \mu$ with $Y \sim \nu$ (section 2.2) in $L_{\mathrm{BMM}}$, and $G_\tau(X)$ with $G_\tau(Y)$ in $L_{\mathrm{CBMM}}$. In this case we recover the RMM loss of Bamberger et al. (2026a) when $Y \mid X = x \sim \mathcal{N}(x, \tau I)$ and $\mathcal{D}(\cdot,\cdot)$ is Frobenius.

## 4.2 LogDet is the natural Bregman divergence for the CDC

We can use $L_{\mathrm{CBMM}}$ with any Bregman divergence $\mathcal{D}$ to learn $G_\tau$. Since $\mathcal{D}$ is defined on the image of $\Phi$, which we presume to be arbitrary, we can ask whether there is a $\mathcal{D}$ that is itself observation-invariant. If $J$ is a $D \times D$ matrix representing a linear change in observation, then $J$ transforms the CDC by conjugation, so an ideal $\mathcal{D}$ would satisfy $\mathcal{D}(JAJ^\top, JBJ^\top) = \mathcal{D}(A, B)$. If $\mathcal{D}(A, B) = \|A - B\|^2$, this holds only when $J$ is orthogonal. For general $J$, this condition specifies $\mathcal{D}$ uniquely.

**Proposition 4.3** (Naturality of the LogDet divergence)**.** *Let $\mathcal{D}$ be a Bregman divergence on the space of $D \times D$ positive definite matrices. Then the following are equivalent:*

1. $\mathcal{D}(JAJ^\top, JBJ^\top) = \mathcal{D}(A, B)$ *for all* $A, B \succ 0$ *and* $J \in \mathrm{GL}(D)$;
2. $\mathcal{D}(A, B) = \mathrm{tr}(B^{-1}A) - \log\det(B^{-1}A) - D$, *up to scale.*

This divergence is called the **LogDet** $\mathcal{D}_{\mathrm{LD}}$, and is a popular loss for matrix regression (Davis et al., 2007; Kulis et al., 2009; James and Stein, 1961). An apparent problem is that $\mathcal{L}_{\mathrm{CBMM}}$ uses rank-1 targets which are not positive definite. However, up to a constant independent of $B$, $\mathcal{D}_{\mathrm{LD}}(A, B) = \mathrm{tr}(B^{-1}A) + \log\det(B)$ is well-defined, so proposition 4.2 still applies. If $\Psi : \mathbb{R}^D \to \mathbb{R}^D$ is an arbitrary injective and differentiable map, representing a non-linear change of observation, then $\Phi$ acts on the CDC by conjugation with the $D \times D$ matrix $D\Psi$. The conjugation-invariance of LogDet then gives asymptotic **non-linear observation-invariance** to $\Psi$ in the limit $\tau \to 0$.

**Corollary 4.4.** *Let* $\Psi : \mathbb{R}^D \to \mathbb{R}^D$ *be a local diffeomorphism at* $x$. *Then, for any* $G^\theta(x) \succ 0$,

$$\mathcal{D}_{\mathrm{LD}}\big(G_{\Psi\circ\Phi}(\Psi(x)), (D_x\Psi)G^\theta(x)(D_x\Psi)^\top\big) = \mathcal{D}_{\mathrm{LD}}\big(G_\Phi(x), G^\theta(x)\big)\,.$$

The proofs are in App. A. These results show that LogDet is the *only observation-invariant loss that permits sample-wise training*. We therefore call LogDet-CBMM **intrinsic geometry matching (IGM)**: both the loss and its target depend only on the process and not on its representation in $\mathbb{R}^D$.

> **Principle 3: LogDet is the natural loss for learning geometry.** It is the only loss that permits sample-wise training and is observation-invariant, so *the training itself* is intrinsic.

**LogDet is just Gaussian NLL.** If $m_\tau^\phi(X)$ models the conditional mean $\mathbb{E}[Y \mid X, \tau]$, then, neglecting terms independent of $\phi$ and $\theta$, we can write the mean-centred CBMM loss as

$$\mathcal{L}_{\mathrm{CBMM}}(\phi, \theta) = \mathbb{E}_{\tau, X, Y|X}\big[\tfrac{1}{2\tau}(Y - m_\tau^\phi(X))^\top G_\tau^\theta(X)^{-1}(Y - m_\tau^\phi(X)) + \log\det G_\tau^\theta(X)\big].$$

Up to adding the Gaussian normalisation constant and dividing by 2, this is the **negative log-likelihood (NLL)** of $Y$ under $\mathcal{N}(m_\tau^\phi(X), 2\tau G_\tau^\theta(X))$[1]. It is minimised by the true mean and covariance (Dawid and Sebastiani, 1999), so can be used to jointly train the parameters $\phi$ and $\theta$.

**Low-rank parametrisation.** The $D \times D$ matrix $G_\tau^\theta$ is too large to materialise in memory for large dimension $D$. If we parametrise $G^\theta = M^\theta(M^\theta)^\top + \epsilon^\theta I_D$, where $M^\theta$ is a $D \times r$ low-rank factor and $\epsilon^\theta$ is a learned scalar, then **LogDet can be evaluated using only the $r \times r$ matrix** $(M^\theta)^\top M^\theta$, allowing easy scaling to high $D$ (App. B). We may prefer a low-rank model and set $\epsilon^\theta = 0$, but this breaks the strict positive definiteness required for LogDet: in this case we use the regularised loss $\mathcal{D}_{\mathrm{LD}}^\lambda(A, B) = \mathcal{D}_{\mathrm{LD}}(A + \lambda I, B + \lambda I)$ for a small $\lambda$, which retains the correct minimiser. For a $d$-dimensional manifold, it suffices to take $r \geq 2d - 1$ (Bamberger et al., 2026a, Thm. 4.4).

### 4.3 Training IGM with data in different spaces

Observation equivariance explicitly relates the geometry of different spaces. We can exploit this to train a model in one space with data from another, potentially unlocking opportunities for scaling data in applications. There are two settings for this, based on the information available.

**Pushing forward the CDC.** Since the CDC pushes forward naturally, this direction is easy. Suppose we have data $X$ and perturbations $Y \mid X$ in $\mathbb{R}^D$, and $\Psi : \mathbb{R}^D \to \mathbb{R}^{D'}$, which could represent an encoder into a latent space. Then we can apply standard IGM to the pushfoward data $\Psi(X)$ and perturbations $\Psi(Y) \mid \Psi(X)$, which is exactly our standard setup of section 3: we regress $G_\tau^\theta$ against rank-1 targets $\Psi(Y) - \Psi(X) \in \mathbb{R}^{D'}$. If $X = \Phi(\alpha)$, for unknown latent $\alpha$, then we approximate the composed CDC $G_{\Psi\circ\Phi}$. Given access to the derivative $D\Psi$, we can also train directly on the conjugated perturbation, giving the rank-1 targets $(D_X\Psi)(Y - X) \in \mathbb{R}^{D'}$.

**Recovering the original CDC.** The harder case is for observed perturbations in $\mathbb{R}^{D'}$, but we want to learn the CDC on $\mathbb{R}^D$. There are several possibilities, but here we assume we have $X \in \mathbb{R}^D$, the map $\Psi$ and its differential $D\Psi$, and perturbations $Z \mid \Psi(X)$ in $\mathbb{R}^{D'}$. We can then formulate the inverse problem for $G_\tau$ by matching the pushed forward CDC against the perturbations in $\mathbb{R}^{D'}$:

$$\mathbb{E}_{\tau, X, Z|\Psi(X)}\big[\mathcal{D}(\tfrac{1}{2\tau}(Z - \Psi(X))(Z - \Psi(X))^\top, (D_X\Psi)G_\tau^\theta(X)(D_X\Psi)^\top)\big]. \tag{11}$$

It may be that $\Psi$ collapses too much information to fully recover $G_\tau$, but may still provide useful additional training when perturbations in $\mathbb{R}^D$ are limited. We could also combine multiple partial data sources to fully determine $G_\tau$: we discuss the formal identifiability criteria in App. C.

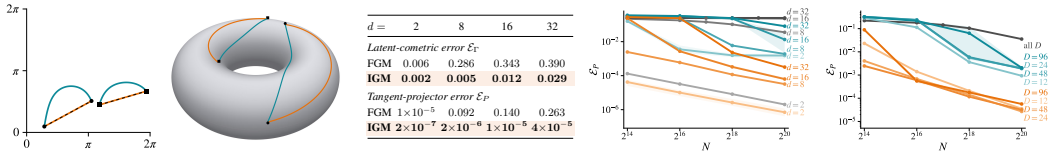
[1] Without a mean model, the Gaussian is $\mathcal{N}(X, 2\tau G_\tau^\theta(X))$.

# 5 EXPERIMENTS

## 5.1 RECOVERING PROCESS GEOMETRY

We test four defining properties of IGM in a controlled setting with known process geometry and observation map: **(i)** recovering the true process geometry under non-isometric observation; **(ii)** the advantage of LogDet over Frobenius loss; **(iii)** scaling laws, in terms of intrinsic and observation dimensions; and **(iv)** recovering intrinsic geometry from an ensemble of partial, rank-deficient views.

We simulate a latent process as Brownian motion on the flat torus $\mathbb{T}^d$, with cometric $I_d$, observed through a curved, non-isometric embedding into $\mathbb{R}^D$ (fig. 2(a)). We train different CDCs on $\mathbb{R}^D$ with different methods: IGM uses the pushforward method with the known differential (section 4.3), FGM (Frobenius geometry matching) denotes the same method but with Frobenius regression not LogDet, and GGM (Gaussian geometry matching) uses ambient Gaussian perturbations with LogDet. Finally, $k$NN estimates a centred, kernel-weighted covariance of clean observations (Jones and Lanners, 2026). Further details are in App. D.1.




(a) (b) (c)

Figure 2: **IGM recovers process geometry with intrinsic-dimensional scaling.** **Orange**: IGM; **teal**: GGM; **grey**: $k$NN. **(a)** Geodesics in latent coordinates (left) and on the observed torus (right): IGM recovers the true geometry, while GGM induces the observation geometry; markers (●, ■) identify endpoint pairs. **(b)** Latent-cometric and tangent-projector errors across intrinsic dimension. **(c)** Tangent-projector error versus sample size, varying intrinsic (left) and observation dimension (right).

**(i)** We compute **geodesics** in latent coordinates for each CDC by minimising the path energy (fig. 2(a)). IGM recovers the true, straight line geodesics (the shortest paths for the process), whereas GGM finds the shortest paths in observation space.

**(ii)** To measure **intrinsic recovery**, we push forward the CDC to the torus through the inverse of the observation, and compare the pointwise Frobenius error to the true cometric $\mathcal{E}_\Gamma$. We compare LogDet and Frobenius regression, for fixed $N = 10^5$ points (fig. 2(b), top). LogDet has a clear advantage, which becomes more pronounced as $d$ increases.

**(iii)** We compare qualitatively different CDCs using the fact that they encode the same tangent space: the top-$d$ eigenvectors at $x \in \mathcal{M} \subset \mathbb{R}^D$ should form an orthogonal projection onto $T_x\mathcal{M}$. We then measure **tangent recovery** by $\mathcal{E}_P$, the squared Frobenius distance to the true rank-$d$ tangent projector. Training models on increasingly more data lets us compare each method's sample efficiency and how it depends on dimension. In fig. 2(c) we vary intrinsic dimension $d$ while fixing ambient dimension $D = 48$ (left), and vary $D$ at fixed $d = 8$ (right), and compare IGM, GGM, and $k$NN. The curves report three-seed means of $\mathcal{E}_P$, and bands show the range. Both neural methods outperform the kernel, as in Bamberger et al. (2026a). However, while GGM infers the tangents from ambient noise, IGM is trained only on intrinsic perturbations which are tangential, giving a low-dimensional signal. This makes IGM much more sample efficient as dimension increases. The effect is especially pronounced when we vary only $D$ (right), where IGM is essentially unaffected.

**(iv)** Following section 4.3, we evaluate whether multiple, individually rank-deficient observations can be combined to fully recover the intrinsic CDC. We use *winding characters* $\psi_{\mathbf{m}}(x) = (\cos\langle \mathbf{m}, x\rangle, \sin\langle \mathbf{m}, x\rangle)$, which are observations $\mathbb{T}^2 \to \mathbb{R}^2$, and each revealing only the rank-one quadratic form $\mathbf{m}^\top \Gamma \mathbf{m}$. As we discuss in App. C, the true CDC is identifiable precisely when $\{\mathbf{m}_o \mathbf{m}_o^\top\}_{o=1}^O$ spans $\mathrm{Sym}(\mathbb{R}^d)$. In this case, the critical threshold is $O^\star = 3$, and we observe this empirically in fig. 3.

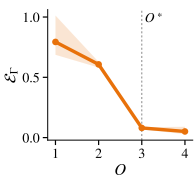
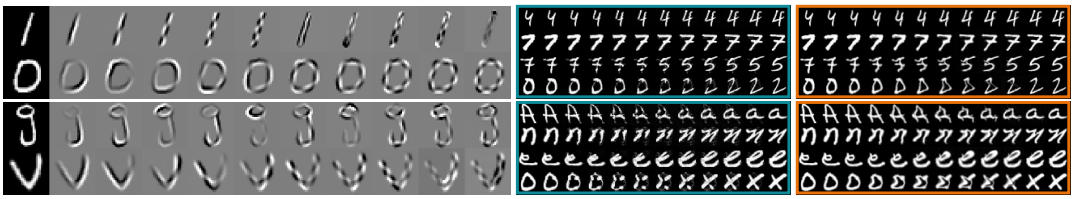



Figure 3: Error vs number of views.

## 5.2 IMAGE GEOMETRY THROUGH AUGMENTATION

When natural process transitions are unavailable, we can simulate them with *data augmentation*. This encompasses Gaussian noise models, where the prior contains no geometric information, so

we now investigate whether semantically meaningful augmentations $Y \sim P^\tau(X, \cdot)$ induce a better geometry. We test whether **(i)** IGM on principled augmentations has better downstream performance than GGM, **(ii)** this process geometry transfers across data distributions without retraining. We set $P^\tau$ to be a smooth spatial deformation in the plane, known as *metamorphosis* (Holm et al., 2008; Younes, 2010; François et al., 2021). IGM learns the uncentred forward field $G_\tau$ from these transitions, whereas GGM learns a mean-centred backward field from isotropic Gaussian perturbations. Both covariance estimators use the same U-Net backbone and rank-100 factorization with a learned isotropic floor. Full details are in App. D.2.



**(a)** **(b)** **(c)**

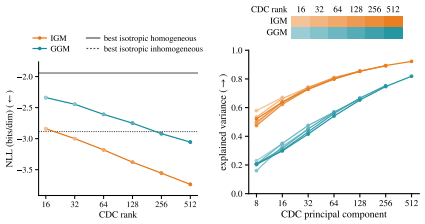
Figure 4: **Process geometry from image transitions.** **(a)** Leading CDC eigenvectors: metamorphic augmentations induce semantically meaningful tangent vectors. **(b)** Interpolating paths with GGM (observation-space geometry). **(c)** Interpolating paths with IGM (intrinsic metamorphic geometry). Models trained on MNIST (top row) are transferred to EMNIST (bottom row) without fine-tuning. Additional eigenvectors across scales and interpolation paths are shown in figs. 8 and 9, respectively.

**(i)** We can visualise the learned CDC geometry from its leading eigenvectors, which generate the tangent space at each point. These learned vectors are highly interpretable, and describe increasingly complex deformations. We apply the CDC to compute interpolating paths between images using *Riemannian optimisation*, taking gradient-descent steps from a source image $x_0$ down the target potential $\frac{1}{2}\|x - x_1\|_2^2$. This is a challenging problem in pixel space, and existing approaches typically use a per-step denoiser to improve path quality (Kharitenko et al., 2026; Kawasaki-Borruat et al., 2026). We evaluate without these tricks, and find that IGM is sufficient for smooth, semantic transformations fig. 4**(c)** while GGM produces irregular paths with pixel-level artifacts fig. 4**(b)**.

**(ii)** We perform the same evaluations on out-of-distribution EMNIST letters without fine-tuning. The GGM interpolating paths lose quality on this unseen data, while IGM has no visible degradation and retains meaningful tangent vectors. This suggests that IGM geometry is *portable*, and can robustly generalise beyond the training distribution.

### 5.3 Natural image geometry from videos

We do have access to *natural image perturbations* from video, where the jumps between time steps represent an intrinsic sample from an underlying process. We train IGM with a low-rank UNet backbone on $64 \times 64$ videos from CelebV-HQ (Zhu et al., 2022), and compare with GGM. We test whether **(i)** IGM with natural perturbations recovers the intrinsic process geometry, **(ii)** how efficiently IGM and GMM learn the tangent spaces of the data. Full details are in App. D.3, and we visualise the learned tangent vectors in fig. 11 and fig. 12.




**(a)** **(b)**

Figure 5: **Natural image geometry from video time steps.** **(a)** NLL (bits/dim) of intrinsic IGM and observation-space GMM, across model rank. **(b)** Explained variance of CDC eigenvectors.

**(i)** We can measure **intrinsic geometry recovery** by interpreting the learned CDC as a Gaussian model and evaluating the negative log-likelihood (NLL) of held-out test transition pairs (fig. 5**(a)**). Since IGM is trained directly on the process, it recovers the true geometry substantially more accurately at each rank, and exhibits a marginally better scaling law as rank increases.

**(ii)** As in the synthetic experiments, we can compare different CDCs by their **tangent space recovery**. For a rank-$r$ model, at each dimension $k \leq r$, we can compute the *explained variance* of the first $k$ eigenvectors. For test transition pairs $(x_t, x_{t+\tau})$, we can measure the fraction of the variance of $x_{t+\tau} - x_t$ retained after projection into the first $k$ CDC principal components, and plot the profile over $k$ in (fig. 5**(b)**). IGM recovers the true tangent spaces in far fewer dimensions than GMM.

### 5.4 EQUIPPING LATENT REPRESENTATIONS WITH PROCESS GEOMETRY

Representation learning is usually focused on achieving a target latent distribution, but this is not always a good proxy for the geometry of the underlying process (Arvanitidis et al., 2017). IGM addresses this mismatch by learning the correct intrinsic geometry in the latent space. Using molecular dynamics in a frozen SLAE representation (Chen et al., 2025), we test whether **(i)** IGM recovers the process geometry in latent coordinates, **(ii)** this geometry improves held-out transition paths, and **(iii)** observation equivariance transports this geometry to physically meaningful molecular motion.

We use 200 dynamic importance sampling trajectories of adenylate kinase (Perilla et al., 2011; Beckstein et al., 2009) and a frozen SLAE encoder $\Phi$ (Chen et al., 2025) which maps each molecular structure $X_t$ to 214 residue-wise embeddings with 128 channels, so $Y_t = \Phi(X_t) \in \mathbb{R}^{214\times128}$. Following section 4.3, we apply IGM directly to the encoded transitions $(Y_t, Y_{t+\tau})$, learning the conditional mean $m_\tau^\phi$ and centred forward field $G_\tau^\theta$. Full details are in App. D.4.

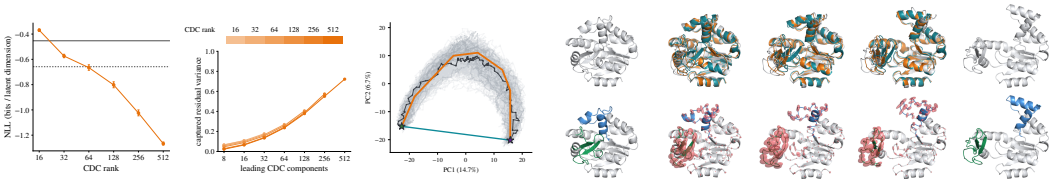



Figure 6: **Process geometry in molecular representations.** **(a)** Held-out NLL and explained variance across CDC rank. **(b)** Held-out MD transition, IGM bridge, and linear interpolation (LERP) between fixed endpoints in latent PCA. **(c)** Corresponding decoded paths (top) and learned CDC pushed forward to per-residue molecular mobility (bottom). **Orange**: IGM; **teal**: LERP; **grey**: held-out MD. Results for all 20 test trajectories are shown in fig. 15; additional decoded paths and mobility ellipsoids are shown in figs. 13 and 14.

**(i)** As in section 5.3, we measure **intrinsic geometry recovery** with NLL and explained variance of the CDC on test trajectories (fig. 6**(a)**). NLL decreases consistently with rank and surpasses the best isotropic baseline, while increasing rank from 16 to 512 raises explained residual variance to 72.2%. Together, these results reveal substantial anisotropic structure in the molecular transition process.

| *Fréchet distance* |
|---|
| LERP $2.00 \pm 0.14$ Å |
| **IGM** $\mathbf{1.40 \pm 0.16}$ Å |

**(ii)** We test whether the learned geometry improves **transition paths** between observed molecular states. We construct an IGM bridge by conditioning the learned transition model on the observed endpoints, using its conditional mean and CDC along the predicted path (App. D.4). Its conditional mean defines the IGM bridge. Figure 6 shows one such transition in latent PCA **(b)** and as decoded structures **(c,** top). We compare each decoded path with the corresponding held-out MD trajectory using discrete Fréchet distance with C$\alpha$ Kabsch-RMSD as ground distance (Seyler et al., 2015). IGM improves on latent linear interpolation for all 20 test trajectories, by 30.1% on average.

**(iii)** We use observation equivariance to push forward the learned CDC through the frozen SLAE decoder $\Psi$ and visualise **molecular mobility** as per-residue displacement ellipsoids (fig. 6**(c,** bottom)). The predicted mobility concentrates in the **LID** and **NMP** domains relative to the **CORE**, consistent with the known open–closed motion of AdK (Beckstein et al., 2009). Thus, the learned latent process geometry can be transported back to reveal physically meaningful molecular motion.

## 6 CONCLUSION

We introduced **Intrinsic Geometry Matching (IGM)**: a scalable approach to learning the intrinsic geometry of data, thereby disentangling the underlying *process geometry* from the distortion of the observation. Our insight is that this geometry can be described with *diffusion geometry* through its carré du champ (CDC) operator, and the CDC can be learned from intrinsic perturbations of the data, even when embedded non-isometrically. This CDC is *observation-equivariant*, transforming naturally between different measurements, and we identify the LogDet divergence as the unique observation-invariant loss for learning it. We can exploit this equivariance to train a model in one space with data from another, potentially unlocking opportunities for scaling data in applications. In each case, it outperformed the Gaussian (observation-dependent) and graph-based estimators. We showed that IGM can induce geometry in latent spaces by learning the process geometry underlying molecular dynamics, recovering accurate trajectories and mobility fields.

### AI USE STATEMENT

Generative AI was used for code development, debugging, and refining figures, and providing feedback on the final manuscript. The authors take full responsibility for the final content of the paper.

### REPRODUCIBILITY STATEMENT

The full details to reproduce experiments are in the Appendix, and we will make the code publicly available upon acceptance.

## REFERENCES


Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. *arXiv preprint arXiv:1710.11379*, 2017.

Dominique Bakry and Michel Émery. Diffusions hypercontractives. In *Séminaire de Probabilités XIX 1983/84: Proceedings*, pages 177–206. Springer, 1985.

Dominique Bakry, Ivan Gentil, and Michel Ledoux. *Analysis and Geometry of Markov Diffusion Operators*, volume 348. Springer Science & Business Media, 2013.

Jacob Bamberger, Adam Gosztolai, Pierre Vandergheynst, Michael Bronstein, and Iolo Jones. Riemannian metric matching for scalable geometric modeling of distributions, 2026a. URL https://arxiv.org/abs/2606.14334.

Jacob Bamberger, Iolo Jones, Dennis Duncan, Michael Bronstein, Pierre Vandergheynst, and Adam Gosztolai. Carré du champ flow matching: better quality-generalisation tradeoff in generative models. In *International Conference on Learning Representations*, volume 2026, pages 135754–135779, 2026b.

Arindam Banerjee, Xin Guo, and Hui Wang. On the optimality of conditional expectation as a bregman predictor. *IEEE Transactions on Information Theory*, 51(7):2664–2669, 2005. doi: 10.1109/TIT.2005.850145.

Oliver Beckstein, Elizabeth J. Denning, Juan R. Perilla, and Thomas B. Woolf. Zipping and unzipping of adenylate kinase: Atomistic insights into the ensemble of open–closed transitions. *Journal of Molecular Biology*, 394(1):160–176, 2009. doi: 10.1016/j.jmb.2009.09.009.

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. arxiv 2012. *arXiv preprint arXiv:1206.5538*, 10, 2012.

Tyrus Berry and Dimitrios Giannakis. Spectral exterior calculus. *Communications on Pure and Applied Mathematics*, 73(4):689–770, 2020.

Yilin Chen, Tianyu Lu, Cizhang Zhao, Hannah K. Wayment-Steele, and Po-Ssu Huang. Slae: Strictly local all-atom environment for protein representation. *bioRxiv*, October 2025. doi: 10.1101/2025.10.03.680398. URL https://www.biorxiv.org/content/10.1101/2025.10.03.680398v1. Preprint.

Ronald R. Coifman and Stéphane Lafon. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006. ISSN 1063-5203. doi: https://doi.org/10.1016/j.acha.2006.04.006. URL https://www.sciencedirect.com/science/article/pii/S1063520306000546. Special Issue: Diffusion Maps and Wavelets.

Jason V Davis, Brian Kulis, Prateek Jain, Suvrit Sra, and Inderjit S Dhillon. Information-theoretic metric learning. In *Proceedings of the 24th international conference on Machine learning*, pages 209–216, 2007.

A Philip Dawid and Paola Sebastiani. Coherent dispersion criteria for optimal experimental design. *Annals of Statistics*, pages 65–81, 1999.

Anton François, Pietro Gori, and Joan Glaunès. Metamorphic image registration using a semi-lagrangian scheme, 2021. URL https://arxiv.org/abs/2106.08817.

Søren Hauberg, Oren Freifeld, and Michael Black. A geometric take on metric learning. *Advances in Neural Information Processing Systems*, 25, 2012.

Darryl D. Holm, Alain Trouve, and Laurent Younes. The euler-poincare theory of metamorphosis, 2008. URL https://arxiv.org/abs/0806.0870.

W. James and Charles Stein. Estimation with quadratic loss. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 361–379. University of California Press, Berkeley, Calif., 1961.

Iolo Jones. Diffusion geometry, 2024. URL https://arxiv.org/abs/2405.10858.

Iolo Jones. Manifold diffusion geometry: Curvature, tangent spaces, and dimension. *SIAM Journal on Mathematics of Data Science*, 8(2):257–277, 2026.

Iolo Jones and David Lanners. Computing diffusion geometry, 2026. URL https://arxiv.org/abs/2602.06006.

Victor Kawasaki-Borruat, Clara Grotehans, Pierre Vandergheynst, and Adam Gosztolai. Diffusion processes on implicit manifolds. *arXiv preprint arXiv:2604.07213*, 2026.

Andrey Kharitenko, Zebang Shen, Riccardo de Santi, Niao He, and Florian Doerfler. Landing with the score: Riemannian optimization through denoising, 2026. URL https://arxiv.org/abs/2509.23357.

Brian Kulis, Mátyás A. Sustik, and Inderjit S. Dhillon. Low-rank kernel learning with bregman matrix divergences. *Journal of Machine Learning Research*, 10:341–376, 2009.

Hiroshi Kunita. Absolute continuity of markov processes and generators. *Nagoya Mathematical Journal*, 36:1–26, 1969.

Guy Lebanon. Learning riemannian metrics. *arXiv preprint arXiv:1212.2474*, 2012.

Hila Manor and Tomer Michaeli. On the posterior distribution in denoising: Application to uncertainty quantification. In *The Twelfth International Conference on Learning Representations*, 2024.

Chenlin Meng, Yang Song, Wenzhe Li, and Stefano Ermon. Estimating high order gradients of the data distribution by denoising. In *Advances in Neural Information Processing Systems*, volume 34, pages 25359–25369, 2021.

Juan R. Perilla, Oliver Beckstein, Elizabeth J. Denning, and Thomas B. Woolf. Computing ensembles of transitions from stable states: Dynamic importance sampling. *Journal of Computational Chemistry*, 32(2):196–209, 2011. doi: 10.1002/jcc.21564.

Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

Sean L. Seyler, Avishek Kumar, M. F. Thorpe, and Oliver Beckstein. Path similarity analysis: A method for quantifying macromolecular pathways. *PLOS Computational Biology*, 11(10): 1–37, 10 2015. doi: 10.1371/journal.pcbi.1004568. URL https://doi.org/10.1371/journal.pcbi.1004568.

Amit Singer and Ronald R Coifman. Non-linear independent component analysis with diffusion maps. *Applied and Computational Harmonic Analysis*, 25(2):226–239, 2008.

Amit Singer, Radek Erban, Ioannis G Kevrekidis, and Ronald R Coifman. Detecting intrinsic slow variables in stochastic dynamical systems by anisotropic diffusion maps. *Proceedings of the National Academy of Sciences*, 106(38):16090–16095, 2009.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in neural information processing systems*, 32, 2019.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Ronen Talmon and Ronald R Coifman. Empirical intrinsic geometry for nonlinear modeling and time series filtering. *Proceedings of the National Academy of Sciences*, 110(31):12535–12540, 2013.

Joshua B Tenenbaum, Vin de Silva, and John C Langford. A global geometric framework for nonlinear dimensionality reduction. *science*, 290(5500):2319–2323, 2000.

Laurent Younes. *Shapes and Diffeomorphisms*. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. ISBN 978-3-642-12054-1. doi: 10.1007/978-3-642-12055-8.

Hao Zhu, Wayne Wu, Wentao Zhu, Liming Jiang, Siwei Tang, Li Zhang, Ziwei Liu, and Chen Change Loy. CelebV-HQ: A large-scale video facial attributes dataset. In *ECCV*, 2022.

## APPENDIX CONTENTS

Notation . . . 13
A. Naturality of the LogDet divergence . . . 13
B. Low-rank loss derivation . . . 15
C. Identifiability from partial observations . . . 16
**D. Experimental details** . . . 17
D.1 Recovering process geometry . . . 17
D.2 Image geometry through augmentation . . . 18
D.3 Natural image geometry from videos . . . 19
D.4 Equipping latent representations with process geometry . . . 24

## NOTATION

| Symbol | Meaning |
|---|---|
| $\mathcal{X}, \mu$ | Latent state space and sampling measure. |
| $\alpha, \Phi, X$ | Latent state, observation map, and observation $X = \Phi(\alpha)$. |
| $d, D, r$ | Intrinsic dimension, observation dimension, and number of columns in the low-rank factor. |
| $\tau$ | Transition lag or perturbation scale. |
| $\mathcal{L}, \Gamma$ | Diffusion generator and carré du champ operator. |
| $g, g^{-1}$ | Riemannian metric and cometric. |
| $D_x\Psi$ | Differential of $\Psi$ at $x$. |
| $\Phi^*, \Phi_\sharp$ | Pullback of functions and pushforward of measures or the CDC, as appropriate. |
| $G_\Phi$ | Coordinate matrix of the observed infinitesimal CDC. |
| $G_\tau$ | Conditional second moment of the observed increment, divided by $2\tau$. |
| $G_{\tau,\mathrm{Cov}}$ | Conditional covariance of the successor, divided by $2\tau$. |
| $m_\tau^\phi, G_\tau^\theta$ | Learned conditional successor mean and learned finite-scale CDC matrix. |
| $M^\theta, \epsilon^\theta$ | Low-rank factor and learned isotropic floor in $G^\theta = M^\theta(M^\theta)^\top + \epsilon^\theta I_D$. |
| $\lambda$ | Fixed additive regularisation of target and model matrices. |
| $\mathcal{D}_\omega, \mathcal{D}_{\mathrm{LD}}$ | Bregman divergence and LogDet divergence. |
| $\mathbb{S}^D, \mathbb{S}^D_+, \mathbb{S}^D_{++}$ | Symmetric, positive-semidefinite, and positive-definite $D \times D$ matrices. |

## A NATURALITY OF THE LOGDET DIVERGENCE

We prove that congruence invariance characterises the LogDet divergence, up to a positive scale, within the class of Bregman divergences considered in Proposition 4.3. The argument first identifies the local Hessian form of a Bregman divergence and classifies congruence-invariant bilinear forms. Requiring such a form to be a Hessian then determines the LogDet generator up to affine terms. Naturality under nonlinear changes of observation follows pointwise from observation equivariance.

**Lemma A.1** (Local Bregman geometry)**.** *Let $\mathcal{D}_\omega$ be as in Definition 4.1, with $\omega \in C^2(\mathcal{A})$. For $B \in \mathrm{int}(\mathcal{A})$,*

$$\mathcal{D}_\omega(B + \epsilon H, B) = \frac{\epsilon^2}{2}\, g_B^\omega(H, H) + o(\epsilon^2), \qquad g_B^\omega(H, K) := \nabla^2\omega(B)[H, K].$$

*Thus $\mathcal{D}_\omega$ locally induces the Hessian metric $g^\omega$.*

*Proof.* Taylor expansion gives

$$\omega(B + \epsilon H) = \omega(B) + \epsilon\langle\nabla\omega(B), H\rangle_{\mathrm{F}} + \frac{\epsilon^2}{2}\nabla^2\omega(B)[H, H] + o(\epsilon^2).$$

Substitution into Definition 4.1 cancels the zeroth- and first-order terms, yielding the claim. □

**Lemma A.2** (Congruence-invariant bilinear forms)**.** *Let $g_B$ be a symmetric bilinear form on $\mathbb{S}^D$, depending on $B \in \mathbb{S}^D_{++}$, and satisfying*

$$g_{JBJ^\top}(JHJ^\top, JKJ^\top) = g_B(H, K)$$

*for all $B \succ 0$, $H, K \in \mathbb{S}^D$, and $J \in \mathrm{GL}(D)$. Then, for $D \geq 2$,*

$$g_B(H, K) = a\ \mathrm{tr}(B^{-1}HB^{-1}K) + b\ \mathrm{tr}(B^{-1}H)\,\mathrm{tr}(B^{-1}K)$$

*for some $a, b \in \mathbb{R}$.*

*Proof.* The congruence action of $\mathrm{GL}(D)$ on $\mathbb{S}^D_{++}$ is transitive, since $B^{-1/2}BB^{-1/2} = I$ for every $B \succ 0$. Hence $g$ is determined by its value at $I$. Since the stabilizer of $I$ is $O(D)$,

$$g_I(QHQ^\top, QKQ^\top) = g_I(H, K), \qquad Q \in O(D).$$

Under orthogonal conjugation,

$$\mathbb{S}^D = \mathbb{R}I \oplus \mathbb{S}^D_0, \qquad \mathbb{S}^D_0 := \{H \in \mathbb{S}^D : \mathrm{tr}(H) = 0\},$$

and the action on $\mathbb{S}^D_0$ is irreducible. Therefore every $O(D)$-invariant symmetric bilinear form has the form

$$g_I(H, K) = a\ \mathrm{tr}(HK) + b\ \mathrm{tr}(H)\,\mathrm{tr}(K).$$

Applying congruence invariance with $J = B^{-1/2}$ gives

$$g_B(H, K) = g_I\left(B^{-1/2}HB^{-1/2}, B^{-1/2}KB^{-1/2}\right).$$

Substituting the expression for $g_I$ and using cyclicity of the trace yields

$$g_B(H, K) = a\ \mathrm{tr}(B^{-1}HB^{-1}K) + b\ \mathrm{tr}(B^{-1}H)\,\mathrm{tr}(B^{-1}K),$$

as claimed. □

**Proposition 4.3** (Naturality of the LogDet divergence)**.** *Let $\mathcal{D}$ be a Bregman divergence on the space of $D \times D$ positive definite matrices. Then the following are equivalent:*

1. *$\mathcal{D}(JAJ^\top, JBJ^\top) = \mathcal{D}(A, B)$ for all $A, B \succ 0$ and $J \in \mathrm{GL}(D)$;*
2. *$\mathcal{D}(A, B) = \mathrm{tr}(B^{-1}A) - \log\det(B^{-1}A) - D$, up to scale.*

*Proof.* Assume first that $\mathcal{D}_\omega$ is congruence invariant. By Lemma A.1, its Hessian form satisfies

$$g^\omega_{JBJ^\top}(JHJ^\top, JHJ^\top) = g^\omega_B(H, H).$$

By the polarization identity $g^\omega_B(H, K) = \frac{1}{2}\big[g^\omega_B(H + K, H + K) - g^\omega_B(H, H) - g^\omega_B(K, K)\big]$, invariance of the quadratic form implies invariance of the full bilinear form. Thus $g^\omega$ is congruence invariant, and Lemma A.2 gives, for $D \geq 2$,

$$g^\omega_B(H, K) = a\ \mathrm{tr}(B^{-1}HB^{-1}K) + b\ \mathrm{tr}(B^{-1}H)\,\mathrm{tr}(B^{-1}K). \tag{12}$$

It remains to impose that $g^\omega$ is a Hessian metric. Since $g^\omega = \nabla^2\omega$, we have $Dg^\omega_B[L](H, K) = \nabla^3\omega(B)[L, H, K]$, which is symmetric in $L, H, K$. Using $D(B^{-1})[L] = -B^{-1}LB^{-1}$, differentiation of equation 12 at $B = I$ in direction $L$ gives

$$\begin{aligned} Dg^\omega_I[L](H, K) = &-a\big[\mathrm{tr}(LHK) + \mathrm{tr}(HLK)\big] \\ &-b\big[\mathrm{tr}(LH)\,\mathrm{tr}(K) + \mathrm{tr}(H)\,\mathrm{tr}(LK)\big]. \end{aligned}$$

Exchanging $L$ and $H$ and using the symmetry of the third derivative therefore yields

$$b\Big[\mathrm{tr}(H)\,\mathrm{tr}(LK) - \mathrm{tr}(L)\,\mathrm{tr}(HK)\Big] = 0.$$

Taking $H = I$ and $K = L$, with $L \neq 0$ traceless, gives $b\,D\ \mathrm{tr}(L^2) = 0$, hence $b = 0$. Therefore

$$\nabla^2\omega(B)[H, K] = a\ \mathrm{tr}(B^{-1}HB^{-1}K) = a\,\nabla^2(-\log\det B)[H, K].$$

Thus $\omega(B) + a\log\det B$ has vanishing Hessian and is affine:

$$\omega(B) = -a\log\det B + \langle C, B\rangle + c_0.$$

Affine terms do not affect Bregman divergences, while strict convexity implies $a > 0$. Consequently,

$$\mathcal{D}_\omega(A, B) = a\Big[\operatorname{tr}(B^{-1}A) - \log\det(B^{-1}A) - D\Big].$$

For $D = 1$, congruence invariance reduces to scale invariance. Applied to the local Hessian form, this gives $\omega''(x) = a/x^2$, and the same conclusion follows.

Conversely, $(JBJ^\top)^{-1}(JAJ^\top) = J^{-\top}(B^{-1}A)J^\top$ is similar to $B^{-1}A$. Since trace and determinant are invariant under similarity,

$$\mathcal{D}_{\mathrm{LD}}(JAJ^\top, JBJ^\top) = \mathcal{D}_{\mathrm{LD}}(A, B),$$

which completes the proof. □

**Corollary 4.4.** *Let $\Psi : \mathbb{R}^D \to \mathbb{R}^D$ be a local diffeomorphism at $x$. Then, for any $G^\theta(x) \succ 0$,*

$$\mathcal{D}_{\mathrm{LD}}\big(G_{\Psi\circ\Phi}(\Psi(x)), (D_x\Psi)G^\theta(x)(D_x\Psi)^\top\big) = \mathcal{D}_{\mathrm{LD}}\big(G_\Phi(x), G^\theta(x)\big)\,.$$

*Proof.* By observation equivariance (equation 7),

$$G_{\Psi\circ\Phi}(\Psi(x)) = (D_x\Psi)G_\Phi(x)(D_x\Psi)^\top.$$

Since $\Psi$ is a local diffeomorphism at $x$, $D_x\Psi \in \mathrm{GL}(D)$. The result then follows directly from the congruence invariance of $\mathcal{D}_{\mathrm{LD}}$ in Proposition 4.3. □

## B Low-rank loss derivation

We derive the low-rank form of the regularised LogDet loss in Section 4.1. For $X, Y \in \mathbb{R}^D$ and fixed $\tau > 0$, define

$$e = Y - m_\tau^\phi(X), \qquad A_\phi = \frac{ee^\top}{2\tau}.$$

The original CBMM target is recovered by setting $e = Y - X$. When the conditional mean is pretrained and frozen, as in the molecular-dynamics experiments, $A_\phi$ is fixed during covariance optimisation.

We parametrise $G_\tau^\theta = M^\theta(M^\theta)^\top + \epsilon^\theta I_D$, where $M^\theta \in \mathbb{R}^{D\times r}$ and $\epsilon^\theta \geq 0$, suppressing dependence on $X$ and $\tau$. For fixed regularisation $\lambda > 0$, we use

$$\mathcal{D}_{\mathrm{LD}}^\lambda(A_\phi, G_\tau^\theta) = \mathcal{D}_{\mathrm{LD}}\big(A_\phi + \lambda I_D, G_\tau^\theta + \lambda I_D\big).$$

Write

$$\eta = \epsilon^\theta + \lambda, \qquad B = G_\tau^\theta + \lambda I_D = M^\theta(M^\theta)^\top + \eta I_D,$$
$$K_\eta = I_r + \eta^{-1}(M^\theta)^\top M^\theta, \qquad z = (M^\theta)^\top e.$$

The regularised covariance score

$$\mathcal{L}_{\mathrm{score}} = \frac{e^\top B^{-1} e}{2\tau} + \lambda \operatorname{tr}(B^{-1}) + \log\det B \tag{13}$$

differs from the regularised LogDet divergence by the target-only terms:

$$\mathcal{D}_{\mathrm{LD}}^\lambda(A_\phi, G_\tau^\theta) = \mathcal{L}_{\mathrm{score}} - \log\det(A_\phi + \lambda I_D) - D.$$

These terms can be discarded when the target is fixed, but not generally when the mean is learned.

The matrix determinant lemma and Woodbury identity give

$$\log\det B = D\log\eta + \log\det K_\eta, \qquad B^{-1} = \eta^{-1}I_D - \eta^{-2}M^\theta K_\eta^{-1}(M^\theta)^\top.$$

In particular,

$$\operatorname{tr}(B^{-1}) = \eta^{-1}\left(D - r + \operatorname{tr}(K_\eta^{-1})\right).$$

Substitution into equation 13 yields

$$\mathcal{L}_{\text{score}} = D\log\eta + \log\det K_\eta + \frac{\|e\|^2}{2\tau\eta} - \frac{z^\top K_\eta^{-1} z}{2\tau\eta^2} + \frac{\lambda}{\eta}\left(D - r + \operatorname{tr}(K_\eta^{-1})\right). \tag{14}$$

Applying the determinant lemma to the target gives the full regularised LogDet divergence:

$$\mathcal{D}^{\lambda}_{\text{LD}}(A_\phi, G^\theta_\tau) = \mathcal{L}_{\text{score}} - D\log\lambda - \log\left(1 + \frac{\|e\|^2}{2\tau\lambda}\right) - D. \tag{15}$$

Both expressions retain all parameter-dependent terms.

For a fixed target and $\epsilon^\theta = 0$, we have $\eta = \lambda$. Writing $K = I_r + \lambda^{-1}(M^\theta)^\top M^\theta$, both objectives reduce, up to parameter-independent terms, to

$$\log\det K + \operatorname{tr}(K^{-1}) - \frac{z^\top K^{-1} z}{2\tau\lambda^2}. \tag{16}$$

This applies to $e = Y - X$ and to $e = Y - m^\phi_\tau(X)$ when the mean is held fixed. For a learned mean with fixed $\eta = \lambda$, the regularised covariance score requires adding $\|e\|^2/(2\tau\lambda)$ to equation 16. The full regularised LogDet divergence requires this residual-norm term and the additional term $-\log(1 + \|e\|^2/(2\tau\lambda))$, up to parameter-independent constants. If the floor is learned, all $\eta$-dependent terms in equation 14 must also be retained. Thus, the regularised covariance score and divergence are equivalent up to parameter-independent terms for a fixed target, but not generally for joint mean–covariance training.

All matrix inverses and determinants needed to evaluate these objectives involve only the $r \times r$ matrix $K_\eta$, while the target determinant reduces to a scalar expression. The remaining computations use $\|e\|^2$ and $z = (M^\theta)^\top e$. Using a Cholesky factorisation and triangular solves gives total cost $O(Dr^2 + r^3)$ and memory $O(Dr + r^2)$ per sample, excluding network evaluation, without materialising a $D \times D$ cometric.

## C Identifiability from partial observations

We characterise when multiple partial observations, as described in Section 4.3, uniquely determine a cometric from its differential pushforwards. For differentiable observation maps $\Psi_o : \mathbb{R}^D \to \mathbb{R}^{D_o}$, $o = 1, \ldots, O$, and a fixed $x \in \mathbb{R}^D$, define

$$\Lambda_x(A) := \left((D_x\Psi_o) A (D_x\Psi_o)^\top\right)_{o=1}^{O}, \qquad A \in \mathbb{S}^D.$$

For a subspace $V \subseteq \mathbb{R}^D$, let

$$\operatorname{Sym}^2(V) := \operatorname{span}\{uv^\top + vu^\top : u, v \in V\} \subseteq \mathbb{S}^D.$$

**Lemma C.1** (Identifiability from partial observations)**.** *The differential observations $\Lambda_x(G_\tau(x))$ uniquely determine every $G_\tau(x) \in \mathbb{S}^D_+$ if and only if*

$$\sum_{o=1}^{O} \operatorname{Sym}^2\left(\operatorname{range}\left((D_x\Psi_o)^\top\right)\right) = \mathbb{S}^D,$$

*equivalently,* $\ker\Lambda_x = \{0\}$.

*Proof.* Write $J_o = D_x\Psi_o$. Under the Frobenius inner product, the adjoint of $A \mapsto J_o A J_o^\top$ is $H_o \mapsto J_o^\top H_o J_o$, whose image is $\operatorname{Sym}^2(\operatorname{range}(J_o^\top))$. Hence

$$(\ker\Lambda_x)^\perp = \operatorname{im}\Lambda_x^* = \sum_{o=1}^{O} \operatorname{Sym}^2(\operatorname{range}(J_o^\top)),$$

which proves the criterion for injectivity on $\mathbb{S}^D$. If $0 \neq A \in \ker\Lambda_x$, then, for sufficiently small $\epsilon > 0$, the distinct matrices $I_D + \epsilon A$ and $I_D - \epsilon A$ are positive definite and have identical observations. Thus injectivity on $\mathbb{S}^D$ is also necessary for identifiability on $\mathbb{S}^D_+$; sufficiency is immediate. □

The condition requires the views to resolve every quadratic interaction, not merely every direction: $\Lambda_x$ must have rank $D(D+1)/2$. For unrestricted ambient cometrics and coordinate projections $\Psi_o(x) = (x_i)_{i \in S_o}$, this means that every coordinate pair $\{i, j\}$, including $i = j$, must occur together in at least one view. We apply this criterion to the partial observations in the synthetic experiments in Section 5.1.

# D Experimental details

## D.1 Recovering process geometry

**Latent process and observation map.** Let $d = 2m$ and $\alpha = (u_1, v_1, \ldots, u_m, v_m) \in \mathbb{T}^d$, with angular coordinates represented in $[0, 2\pi)$. We sample independent latent anchors $\alpha_0 \sim \mathrm{Unif}(\mathbb{T}^d)$ and Brownian transitions

$$\alpha_\tau = \big(\alpha_0 + \sqrt{2\tau}\,\xi\big) \pmod{2\pi}, \qquad \xi \sim \mathcal{N}(0, I_d),$$

where $\xi$ is independent of $\alpha_0$. This process has generator $\Delta = \sum_{i=1}^{d} \partial^2_{\alpha_i}$ and cometric $I_d$.

We observe each two-dimensional factor through the torus-of-revolution embedding

$$\Phi_0(u, v) = \begin{pmatrix} (R + r\cos v)\cos u \\ (R + r\cos v)\sin u \\ r \sin v \end{pmatrix}, \qquad R = 2, \quad r = 1. \tag{17}$$

Concatenating the $m$ blocks gives $\Phi = \Phi_0^{\times m} : \mathbb{T}^d \to \mathbb{R}^H$, where $H = 3d/2$. For each block,

$$(D_{(u,v)}\Phi_0)^\top D_{(u,v)}\Phi_0 = \mathrm{diag}\big((R + r\cos v)^2, r^2\big),$$

so the observation-induced metric differs from the flat process metric.

To vary the observation dimension, we use the isometric lift $\Psi_D(\alpha) = Q_D\Phi(\alpha)$ for $D \geq H$, where $Q_D \in \mathbb{R}^{D \times H}$ satisfies $Q_D^\top Q_D = I_H$. For $D = H$, we set $Q_D = I_H$; for $D > H$, we obtain $Q_D$ by QR factorisation of a fixed Gaussian matrix. The observation-induced metric is preserved:

$$(D_\alpha \Psi_D)^\top D_\alpha \Psi_D = (D_\alpha \Phi)^\top D_\alpha \Phi.$$

**Neural estimators.** All neural CDC estimators use the low-rank parametrisation of App. B with rank $r = 64$. The backbone is a FiLM-conditioned residual MLP with width 512, four residual blocks, and a 128-dimensional sinusoidal embedding of the scale $\tau$. All models are trained with Adam using a learning rate of $2 \times 10^{-4}$ and batch size 1024.

**Evaluation and uncertainty.** Within each setting, all methods are evaluated on the same fixed set of 8192 held-out latent anchors $\alpha \sim \mathrm{Unif}(\mathbb{T}^d)$, excluded from training and model selection. The true tangent projector in observation space is

$$P(\alpha) = (D_\alpha \Psi_D)\big((D_\alpha \Psi_D)^\top D_\alpha \Psi_D\big)^{-1}(D_\alpha \Psi_D)^\top.$$

For each method, let $\widehat{P}(\alpha)$ be the orthogonal projector onto the span of the leading $d$ eigenvectors of its estimated ambient CDC at $\Psi_D(\alpha)$. We measure tangent recovery by

$$\mathcal{E}_P = \mathbb{E}_\alpha \left[ \frac{\|\widehat{P}(\alpha) - P(\alpha)\|_{\mathrm{F}}^2}{2d} \right], \tag{18}$$

which is the average over anchors of the mean squared sine of the principal angles between the estimated and true tangent spaces.

For IGM and FGM, we additionally recover the learned cometric in latent coordinates:

$$G^{\theta,\mathrm{lat}}(\alpha) = (D_\alpha \Psi_D)^\dagger G^\theta(\Psi_D(\alpha))\big((D_\alpha \Psi_D)^\dagger\big)^\top,$$

where $\dagger$ denotes the Moore–Penrose pseudoinverse and the evaluation scale is suppressed in the notation. Since the true latent cometric is $I_d$, we report the relative latent-cometric error

$$\mathcal{E}_\Gamma = \mathbb{E}_\alpha \left[ \frac{\|G^{\theta,\mathrm{lat}}(\alpha) - I_d\|_{\mathrm{F}}}{\sqrt{d}} \right]. \tag{19}$$

We estimate both expectations by averaging over the held-out anchors and report averages across training seeds. For the sample-size experiments, curves show three-seed means and bands span the seed range.

**Geodesic optimisation.** We compute geodesics in latent coordinates using the metric $g^\theta$ obtained from the learned cometric. For fixed endpoints $\alpha_0, \alpha_K \in \mathbb{T}^2$, we optimise $K = 64$ segments by minimising the midpoint Dirichlet energy

$$E(\alpha_0, \dots, \alpha_K) = K \sum_{k=0}^{K-1} \Delta\alpha_k^\top g^\theta\big(\alpha_k + \tfrac{1}{2}\Delta\alpha_k\big)\,\Delta\alpha_k, \qquad \Delta\alpha_k = \mathrm{wrap}(\alpha_{k+1} - \alpha_k) \in [-\pi, \pi)^2. \tag{20}$$

Interior vertices are initialised along the shortest wrapped linear path and optimised for 1200 Adam iterations with learning rate 0.03, cosine-decayed to zero. Both estimators use the same endpoint pairs: $((0.9, 0.3), (3.3, 1.6))$ and $((3.7, 1.4), (6.1, 2.1))$.

**Rank-deficient views.** For **(iv)**, we set $d = 2$ and replace the isotropic Brownian motion above with an anisotropic diffusion on the flat torus:

$$\alpha_\tau = \big(\alpha_0 + \sqrt{2\tau}\, A^{1/2}\xi\big) \pmod{2\pi}, \qquad \xi \sim \mathcal{N}(0, I_2), \qquad A = \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}.$$

This process has cometric $A$, whose off-diagonal entries describe correlated fluctuations of the two angular coordinates. We observe it through winding characters

$$\psi_{\mathbf{m}}(\alpha) = \big(\cos\langle \mathbf{m}, \alpha\rangle, \sin\langle \mathbf{m}, \alpha\rangle\big), \qquad \mathbf{m} \in \mathbb{Z}^2 \setminus \{0\}.$$

Each map records the angle $\langle \mathbf{m}, \alpha\rangle$ modulo $2\pi$, so it is globally non-injective and its differential has rank one. Writing $t_{\mathbf{m}}(\alpha) = (-\sin\langle \mathbf{m}, \alpha\rangle, \cos\langle \mathbf{m}, \alpha\rangle)^\top$, we have $D_\alpha \psi_{\mathbf{m}} = t_{\mathbf{m}}(\alpha)\mathbf{m}^\top$ and

$$(D_\alpha \psi_{\mathbf{m}}) A (D_\alpha \psi_{\mathbf{m}})^\top = (\mathbf{m}^\top A \mathbf{m})\, t_{\mathbf{m}}(\alpha) t_{\mathbf{m}}(\alpha)^\top.$$

Each differential observation therefore reveals only the quadratic measurement $\mathbf{m}^\top A \mathbf{m}$. We use the nested sets

$$\mathcal{M}_1 = \{(1, 0)\}, \qquad \mathcal{M}_2 = \mathcal{M}_1 \cup \{(0, 1)\},$$
$$\mathcal{M}_3 = \mathcal{M}_2 \cup \{(1, 1)\}, \quad \mathcal{M}_4 = \mathcal{M}_3 \cup \{(1, -1)\}.$$

By Lemma C.1, these differential observations identify an unrestricted latent cometric precisely when $\{\mathbf{m}\mathbf{m}^\top : \mathbf{m} \in \mathcal{M}_O\}$ spans $\mathbb{S}^2$. The first two views reveal the diagonal entries but leave the cross term unresolved; the third completes the span, giving the threshold $O^\star = 3$.

For each $O \in \{1, 2, 3, 4\}$, we train IGM on $N = 10^5$ anchors for 200 epochs with four seeds. We use rank $r = 8$, width 512, four residual blocks, batch size 1024, and $\tau \sim \mathrm{Unif}[0.005, 0.1]$, with the debiased LogDet loss and $\lambda = 0.1$. We evaluate at $\tau = 0.005$ on 4000 held-out anchors and report the latent-cometric error against $A$.

### D.2 Image geometry through augmentation

**Data and transition process.** We train on the official MNIST training split using $28 \times 28$ greyscale images. For IGM, given an anchor $X$, we sample image momentum $z = 0.3\varepsilon$, where $\varepsilon \sim \mathcal{N}(0, I)$, and generate a metamorphic transition by $v = -\eta K_\sigma \star (z\nabla X)$, with $\phi_\tau = \exp(-\tau v)$, and $Y = X \circ \phi_\tau + \frac{\tau\kappa}{2}(z + z \circ \phi_\tau)$. Here, $K_\sigma$ is a Gaussian smoothing kernel, $\star$ denotes spatial convolution, and $\phi_\tau$ is the time-$\tau$ flow of the stationary velocity field $-v$. The transition combines spatial deformation with an additive intensity change. We use $\eta = 1.2$, $\kappa = 0.05$, and $\sigma = 2$ pixels with $13 \times 13$ kernel support, sampling $\tau \sim \mathrm{LogUnif}[0.2, 0.4]$ independently per example. Spatial derivatives use centred differences, and generated transitions are not clipped. Examples are shown in Figure 7.

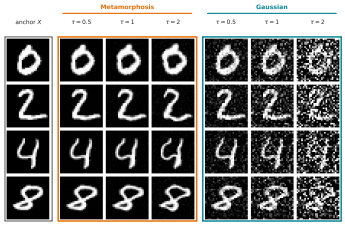



Figure 7: **Sampled image transitions.** Samples at increasing transition scales $\tau \in \{0.5, 1, 2\}$. **Orange**: metamorphosis; **teal**: isotropic Gaussian perturbations.

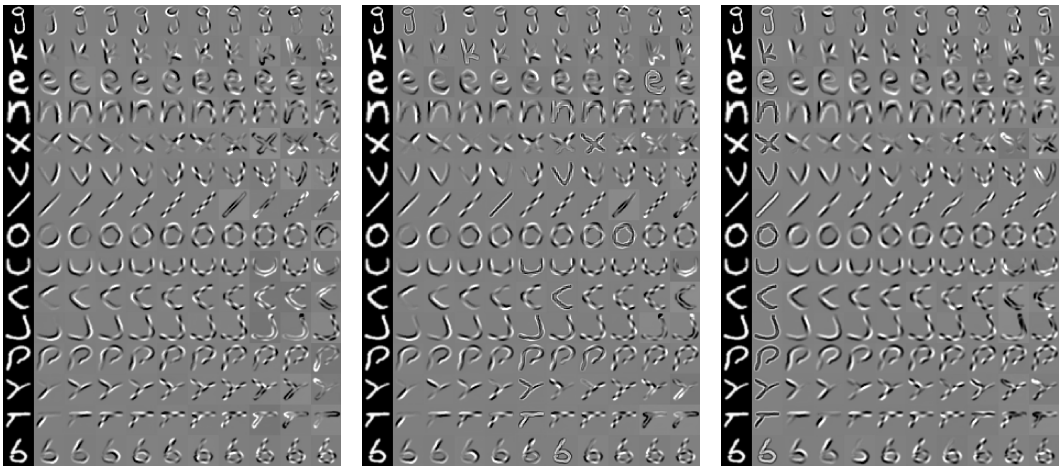
**Field estimation and training.** IGM learns the uncentred forward field from metamorphic transitions, while GGM learns a mean-centred backward field from isotropic Gaussian perturbations. Both CDC estimators use a scale-conditioned U-Net with base width 64, channel multipliers $(1, 2, 2)$, two residual blocks per resolution, and no dropout. Single-head self-attention is applied at downsampling factor 4 and in the middle block. A sinusoidal embedding of $\tau$, followed by a two-layer MLP, conditions each residual block. The CDC head outputs

100 spatial channels, flattened into $U_\theta \in \mathbb{R}^{100\times 784}$, and the cometric includes a learned isotropic floor. We use Adam with learning rate $2 \times 10^{-4}$, batch size 128, no weight decay, and gradient clipping at norm 1. After the fixed validation objective plateaus for 15 epochs, we trigger a 100-epoch cosine decay of the learning rate to zero, with training capped at 200 epochs.

**Target-directed trajectories.** As described in the main text, we compute interpolating paths by Riemannian optimisation, evaluating the learned field at $\tau = 0.3$. For each source–target pair $(x_0, x_1)$, we take normalised Riemannian-gradient steps on the target potential $\frac{1}{2}\|x - x_1\|_2^2$, using the learned low-rank cometric:

$$g_k = x_k - x_1, \qquad d_k = \frac{U_k^\top U_k g_k}{\|U_k g_k\|_2}, \qquad x_{k+1} = x_k - \alpha d_k. \tag{21}$$

Here, $U_k$ is evaluated at $x_k$, and the denominator normalises the Riemannian-gradient direction. We use 8192 updates and calibrate the step size $\alpha$ once per field using the same source–target pairs. The field is re-evaluated at every iterate, with no denoising, projection, or post-step clipping. We apply the MNIST-trained models to EMNIST letters without fine-tuning. Figure 8 shows eigenvectors at additional scales, and Figure 9 shows additional paths.



(a) (b) (c)

Figure 8: **IGM eigenvectors across transition scales.** Leading CDC eigenvectors of the MNIST-trained model, evaluated on unseen EMNIST letters at **(a)** $\tau = 0.05$, **(b)** $\tau = 0.1$, and **(c)** $\tau = 0.2$. Larger scales capture progressively larger, morphology-aligned variations.

### D.3 Natural image geometry from videos

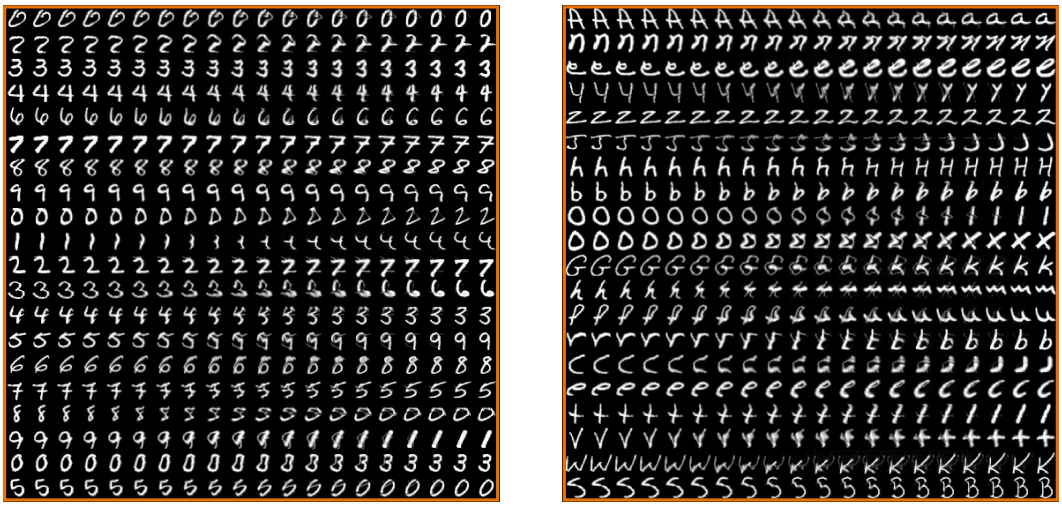
**Data and transition process.** We use the processed CelebV-HQ clips (Zhu et al., 2022), decoding RGB frames with their original presentation timestamps and resizing them to $64\times64$, giving ambient dimension $D = 3 \cdot 64^2 = 12{,}288$. We use a 95%/5% split for training and validation. Excluding clips longer than 20 seconds leaves 33,594 training clips and 1,825 validation clips.

IGM is trained on forward pairs $(X, Y) = (x_t, x_{t+\tau})$ from a single clip, with $\tau \in [0.04, 0.5]$ seconds. We pick a source frame uniformly among those with an eligible successor, sample a lag $\tau$ log-uniformly in the range, and choose the successor closest in log-lag. We condition the network and normalise the loss with the actual timestamp difference in seconds, rather than the requested lag or a frame-index offset. In each training epoch, we sample just one pair per clip, so use a negligible fraction of the available training data even when training for over 1000 epochs.

**Gaussian perturbation baseline.** GGM uses still frames from the same training clips, with one uniformly sampled frame per selected clip. For a clean frame $X$, we generate

$$\widetilde{X} = X + \sqrt{2b}\,\xi, \qquad \xi \sim \mathcal{N}(0, I_D), \qquad b \sim \mathrm{Unif}[4.5 \times 10^{-6}, 1.125], \tag{22}$$

so the corresponding noise standard deviation ranges from 0.003 to 1.5 in $[0, 1]$ pixel units. GGM learns the mean-centred posterior field $\mathrm{Cov}[X \mid \widetilde{X} = \widetilde{x}, b]/(2b)$. The forward Gaussian field would be the constant $I_D$ and would contain no image-dependent geometry.



(a) (b)

Figure 9: **Additional IGM interpolation paths.** Paths between randomly sampled source–target pairs from held-out images in **(a)** MNIST and **(b)** EMNIST. The MNIST-trained model produces smooth, structured transformations on both datasets, without fine-tuning on EMNIST.

**Field estimation and training.** Both estimators use the same scale-conditioned U-Net, with base width 64, channel multipliers $(1, 2, 3, 4)$, two residual blocks per encoder resolution, and three per decoder resolution. Single-head self-attention is applied at downsampling factor 8 and in the middle block. To condition the residual blocks on time, we form a 64-dimensional sinusoidal embedding of the scale (divided by 0.04, which represents one frame at 25 frames per second) and apply a two-layer MLP of width 256 to condition the residual blocks. The output has $3(r + 1)$ spatial channels: one RGB displacement mean $a^\theta_\tau(q)$ and $r$ RGB factor rows, flattened to $F^\theta_\tau(q) \in \mathbb{R}^{r\times D}$. We represent

$$G^\theta_\tau(q) = F^\theta_\tau(q)^\top F^\theta_\tau(q) + \epsilon^\theta I_D. \tag{23}$$

Here $\epsilon^\theta = \mathrm{softplus}(\rho) + 10^{-8}$ is a single learned scalar shared across images and scales, initialized at 0.01. So $r$ is the rank of the structured component, while the complete covariance is full rank. For IGM, $q = X$ and $a^\theta_\tau(X)$ predicts $Y - X$, so the endpoint mean is $m^\theta_\tau(X) = X + a^\theta_\tau(X)$. The mean and factor rows share the U-Net and are learned jointly.

For the objective, we use the $\lambda$-regularised LogDet, which corresponds to Gaussian cross-entropy. To specify the objective, let $d$ be the target displacement, $u$ its training scale ($\tau$ for IGM and $b$ for GGM), and $z = (d - a^\theta_u(q))/\sqrt{2u}$. Writing $C_\lambda = G^\theta_u(q) + \lambda I_D$, we define sample loss as cross-entropy in bits/dimension:

$$\begin{aligned}\ell_\lambda(q, d, u) = \frac{1}{2D\log 2}\Big[&D\log(2\pi) + D\log(2u) + \log\det C_\lambda \\ &+ z^\top C_\lambda^{-1} z + \lambda\,\mathrm{tr}(C_\lambda^{-1})\Big].\end{aligned} \tag{24}$$

The smoothing $\lambda$ is added to both the model CDC and the rank-one target $zz^\top$, producing the trace term in Equation (24). When $\lambda = 0$, this is just Gaussian NLL.

For the rank study in fig. 5, we train $r \in \{16, 32, 64, 128, 256, 512\}$ with $\lambda = 0.01$ for 60,000 optimizer steps. We use Adam with $(\beta_1, \beta_2) = (0.9, 0.999)$, $\epsilon = 10^{-8}$, zero weight decay, and constant learning rate $5 \times 10^{-5}$. Each update uses 384 sampled clips, contributing one real transition per clip for IGM or one corrupted still frame per clip for GGM. Comparisons use the final model after 60,000 steps, rather than selecting a checkpoint by its best score.

**Gaussian NLL and evaluation smoothing.** The score for transition $i$ is the exact negative log-density of $d_i$ under the Gaussian surrogate

$$d_i \mid q_i, \tau_i \;\sim \mathcal{N}\big(a^\theta_{\tau_i}(q_i),\, 2\tau_i G^\theta_{\tau_i}(q_i)\big)\,, \qquad \mathrm{NLL} = \frac{1}{N}\sum_{i=1}^{N} \ell_0(q_i, d_i, \tau_i). \tag{25}$$

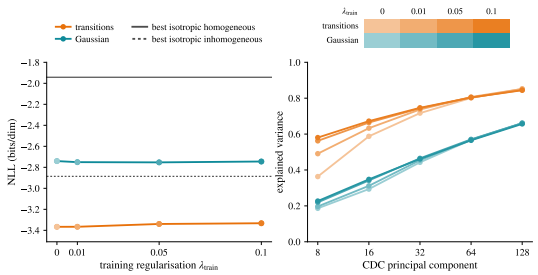



Figure 10: **Training-smoothing ablation on CelebV-HQ.** Left: validation NLL, varying training smoothing $\lambda$ at rank 128. Right: explained transition variance in the leading $k$ CDC directions. **Orange**: IGM; **teal**: GGM. Isotropic NLL references are defined in eq. (27).

This is reported in bits per ambient dimension. It measures mean prediction, covariance calibration, and directional structure together.

**Explained transition variance.** Let $v_{ij}$ be the $j$th unit eigenvector of $F^{\theta}_{\tau_i}(q_i)^{\top} F^{\theta}_{\tau_i}(q_i)$, ordered by decreasing eigenvalue, and $P_{ik} = \sum_{j=1}^{k} v_{ij} v_{ij}^{\top}$. For $k \leq r$, we report the explained variance as a fraction

$$\mathrm{EV}(k) = \frac{\sum_{i=1}^{N} \|P_{ik}(y_i - x_i)\|_2^2 (2\tau_i)^{-1}}{\sum_{i=1}^{N} \|y_i - x_i\|_2^2 (2\tau_i)^{-1}}. \tag{26}$$

We aggregate the normalised energies over all transitions before dividing, rather than averaging per-pair fractions. The statistic measures how much observed motion lies in the leading learned directions. The profile is invariant to positive rescaling of the CDC at a fixed readout and is unaffected by adding an isotropic ridge, which changes eigenvalues but not eigenvectors. We report $k \in \{8, 16, 32, 64, 128, 256, 512\}$ up to the model rank. Small $k$ probes the ordering of useful directions, whereas $k = r$ probes the coverage of the learned span.

**Isotropic reference scores.** We add two isotropic references to the NLL figures, showing the best possible scores attainable by isotropic models, both homogeneous and inhomogeneous. For an isotropic raw-displacement covariance $\sigma_i^2 I_D$, let $w_i = \|y_i - x_i\|_2^2 / D$. The NLL in bits per dimension is $[\log(2\pi) + \log \sigma_i^2 + w_i/\sigma_i^2]/(2 \log 2)$. For the **homogeneous model**, the optimal constant variance is $\bar{w} = N_b^{-1} \sum_i w_i$. The *inhomogeneous model* uses a per-transition scale $\sigma_i^2 = w_i$: it knows each held-out transition's magnitude and is an *oracle reference*, not an independently fitted predictor, but records the best score possibly attainable by an isotropic model. We therefore have

$$\mathrm{NLL}_{\mathrm{hom}} = \tfrac{1}{2} \log_2(2\pi \mathrm{e}\, \bar{w}), \qquad \mathrm{NLL}_{\mathrm{inhom}} = \frac{1}{2N_b} \sum_{i=1}^{N_b} \log_2(2\pi \mathrm{e}\, w_i). \tag{27}$$

**Training-smoothing ablation.** We fix $r = 128$ and vary $\lambda_t \in \{0, 0.01, 0.05, 0.1\}$. As shown in Figure 10, increasing IGM smoothing from 0 to 0.1 raises $\mathrm{EV}(k)$ for small $k$, concentrating more observed motion in the leading directions, while minimally affecting NLL. We visualise the CDC eigenvectors in for $\lambda = 0$ (fig. 11) and $\lambda = 0.1$ (fig. 12), which show the effect of $\lambda$ in reordering the leading CDC directions.

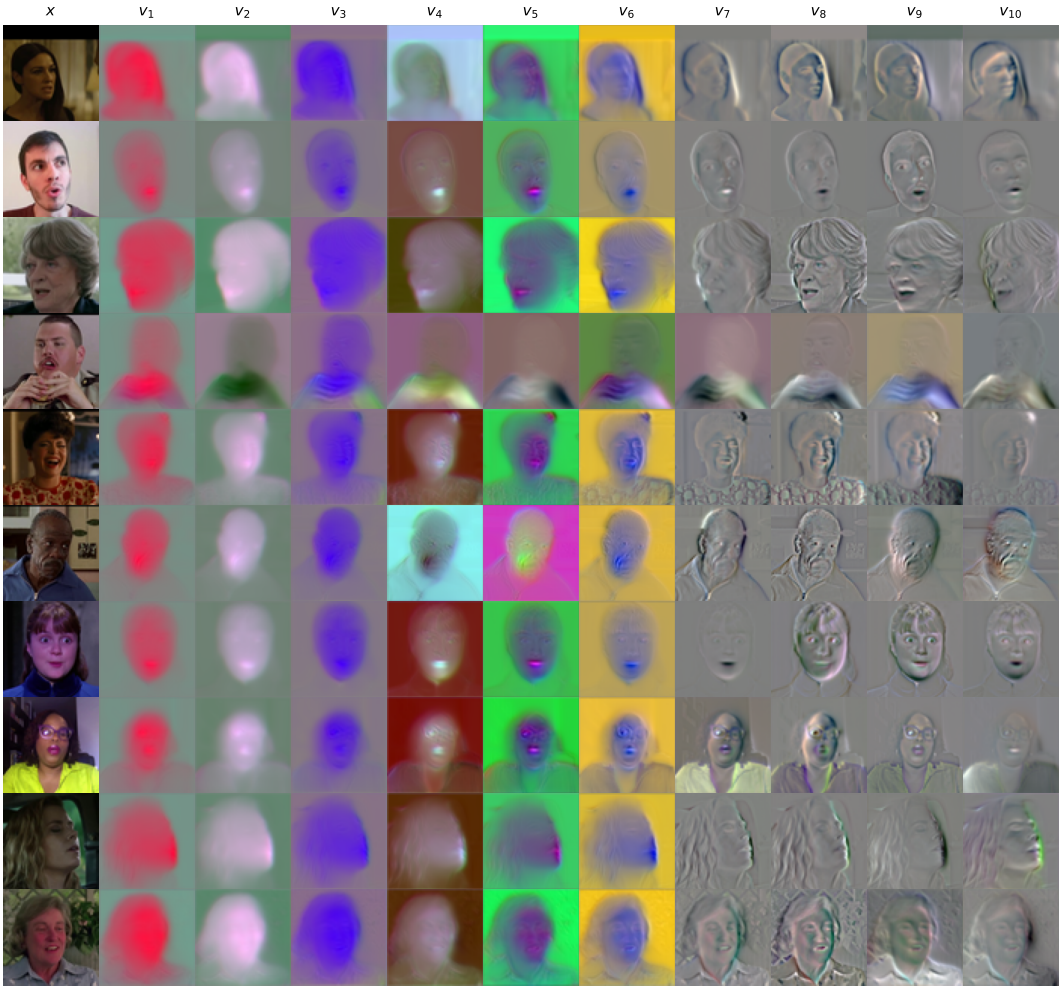



Figure 11: **Intrinsic IGM geometry of images, with unregularised LogDet.** Leading CDC eigenvectors on unseen validation images, with no training smoothing.

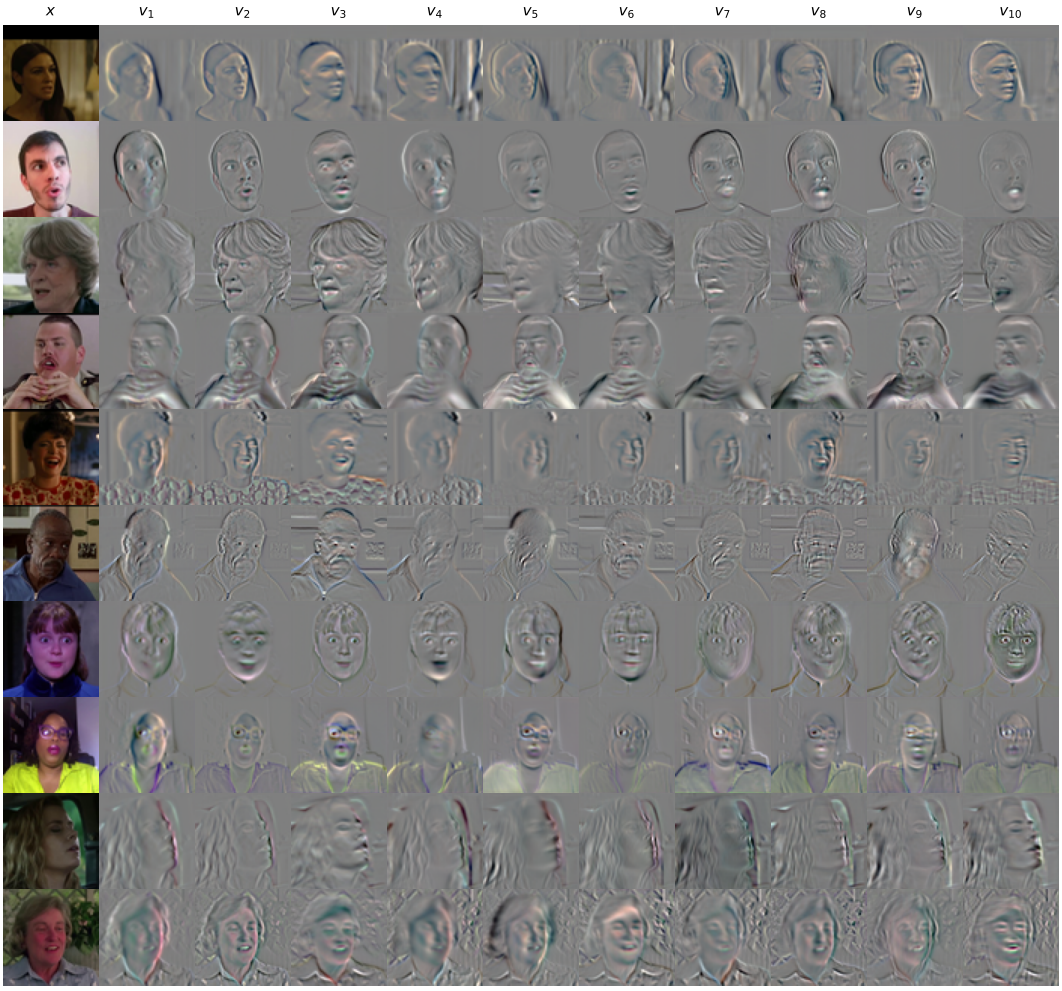



Figure 12: **Intrinsic IGM geometry of images, with LogDet regularisation $\lambda = 0.1$.** Leading CDC eigenvectors on unseen validation images, with training smoothing $\lambda = 0.1$.

### D.4 EQUIPPING LATENT REPRESENTATIONS WITH PROCESS GEOMETRY

**Data and representation.** We use 200 dynamic importance sampling trajectories of adenylate kinase (AdK) (Perilla et al., 2011; Beckstein et al., 2009), comprising 19,691 frames sampled at 1 ps intervals and split by trajectory into 160/20/20 training, validation, and test trajectories. AdK consists of a relatively stable CORE and the mobile **LID** and **NMP** domains, whose relative motion characterises the open–closed transition (Beckstein et al., 2009). Training uses all valid within-trajectory pairs $(X_t, X_{t+\tau})$ at integer lags $\tau \in \{1, \dots, 32\}$ ps, giving 419,136 transitions. We freeze the pretrained SLAE encoder $\Phi$ and decoder $\Psi$ (Chen et al., 2025). SLAE represents each residue by an embedding of its local all-atom environment and decodes the resulting embeddings back to molecular coordinates. The encoder gives $Y_t = \Phi(X_t) \in \mathbb{R}^{214\times128}$, or $D = 27{,}392$ latent coordinates, which we standardise using training-set statistics before fitting the transition models.

**Architecture and training.** The mean and CDC networks preserve the $214 \times 128$ residue structure through residue-shared projections, learned positional embeddings, and residual blocks combining LayerNorm, four-head self-attention, and FiLM-conditioned MLPs with a sinusoidal embedding of $\tau$. The mean network uses width 128 and three residual blocks; the CDC network uses width 256 and four. We first fit the conditional mean $m^\phi_\tau(Y_t) \approx \mathbb{E}[Y_{t+\tau} \mid Y_t]$ with Adam at learning rate $2 \times 10^{-4}$, batch size 64, 200 batches per epoch, zero weight decay, and gradient clipping at norm 1, for at most 400 epochs with patience 20 on validation displacement. We then freeze $m^\phi$ and train CDCs of rank $r \in \{16, 32, 64, 128, 256, 512\}$ on the centred residuals $Y_{t+\tau} - m^\phi_\tau(Y_t)$. This two-stage procedure has the same population optimum as joint Gaussian regression, while allowing the mean to be fitted once and shared across the rank sweep. The CDCs use IGM with low-rank parametrisation $G^\theta_\tau = M^\theta_\tau (M^\theta_\tau)^\top + \epsilon^\theta I$. Adam at learning rate $10^{-4}$, and otherwise the same batch construction, for at most 300 epochs with patience 15. Checkpoints are selected by Gaussian NLL on 4096 validation transitions, and $\epsilon^\theta$ is initialized at $10^{-2}$.

**Held-out transition geometry.** For Figure 6(a), we evaluate all ranks on the same 4096 transitions from the 20 held-out test trajectories. We report Gaussian NLL as in equation 25, scoring the successor state under $\mathcal{N}(m^\phi_\tau(Y_t), 2\tau G^\theta_\tau(Y_t))$ in bits per standardised latent dimension. We compute 95% confidence intervals by resampling the 20 test trajectories 10,000 times, keeping transitions from each trajectory together.

Unlike the uncentred video score in equation 26, we measure residual variability after subtracting the predicted conditional mean. For transition $i$, let

$$\delta_i = \frac{Y_{t_i+\tau_i} - m^\phi_{\tau_i}(Y_{t_i})}{\sqrt{2\tau_i}},$$

and let $P_{ik}$ project onto the leading $k$ eigenvectors of $G^\theta_{\tau_i}(Y_{t_i})$. For $k \le r$, we report

$$\mathrm{EV}_{\mathrm{res}}(k) = \frac{\sum_i \|P_{ik}\delta_i\|_2^2}{\sum_i \|\delta_i\|_2^2}. \tag{28}$$

This measures the fraction of residual variability captured by the leading CDC directions.

**Endpoint-conditioned bridge paths.** For Figure 6(b,c), we condition the learned transition model on the observed endpoints of each held-out trajectory. Starting from the initial latent state $f_0 = Y_0$, we first roll out the learned mean on a 16 ps grid,

$$f_{k+1} = m^\phi_{\tau_k}(f_k),$$

where $\tau_k = 16$ ps except for a shorter final step when required. Along this free-mean path we freeze the predictive covariances $Q_k = 2\tau_k G^\theta_{\tau_k}(f_k)$. These define a time-inhomogeneous additive Gaussian surrogate, whose exact conditional mean given the observed terminal state $Y_T$ is

$$\bar{Y}_j = f_j + \left(\sum_{k<j} Q_k\right)\left(\sum_{k<K} Q_k\right)^{-1}(Y_T - f_K). \tag{29}$$

We solve the corresponding linear system in double precision and decode the resulting bridge with the frozen SLAE decoder. For evaluation, predicted and observed paths are resampled to 80 points by latent arclength and compared by discrete Fréchet distance using C$\alpha$ Kabsch-RMSD as the

ground distance (Seyler et al., 2015). Equation (29) is the exact bridge mean of the frozen Gaussian surrogate, rather than of the nonlinear learned transition process. For the exploratory path analysis, we focus on CDC ranks 32–128 and evaluate eight lags on 20 validation trajectories. Rank 32 at 16 ps has the lowest observed mean Fréchet distance in this subset.

**Transport to molecular coordinates.** For Figure 6(c), bottom, we transport the learned CDC back to molecular coordinates through the frozen SLAE decoder $\Psi$. Let $\Psi$ denote the map from standardised latent coordinates to CORE-aligned C$\alpha$ coordinates, including inverse latent standardisation and rigid alignment. Following the observation-equivariance rule of equation 7, the linearised predictive covariance at latent state $Y$ is

$$\Sigma_X(Y) = \left(D_{m_\tau^\phi(Y)}\Psi\right)\left[2\tau G_\tau^\theta(Y)\right]\left(D_{m_\tau^\phi(Y)}\Psi\right)^\top \in \mathbb{S}_+^{3\cdot 214}. \tag{30}$$

For the main-figure ellipsoids, we use the selected rank-32 CDC at $\tau = 16\,\mathrm{ps}$ and visualise each residue's $3 \times 3$ diagonal block as a one-standard-deviation mobility ellipsoid.

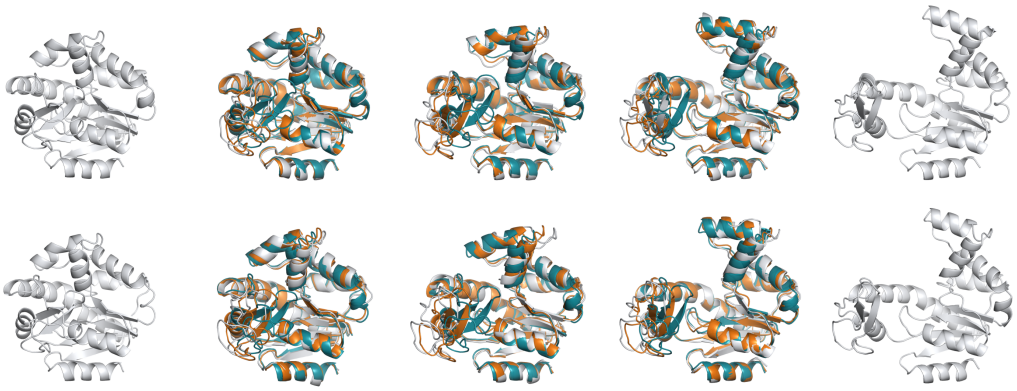


Figure 13: **Decoded endpoint-conditioned paths.** Two held-out MD transitions (trajectories 111 and 145; top and bottom) compared with the decoded IGM bridge and latent linear interpolation (LERP) at matched path progress. Bridges use rank $r = 32$ and lag $\tau = 16\,\mathrm{ps}$. **Orange**: IGM; **teal**: LERP; **grey**: held-out MD.

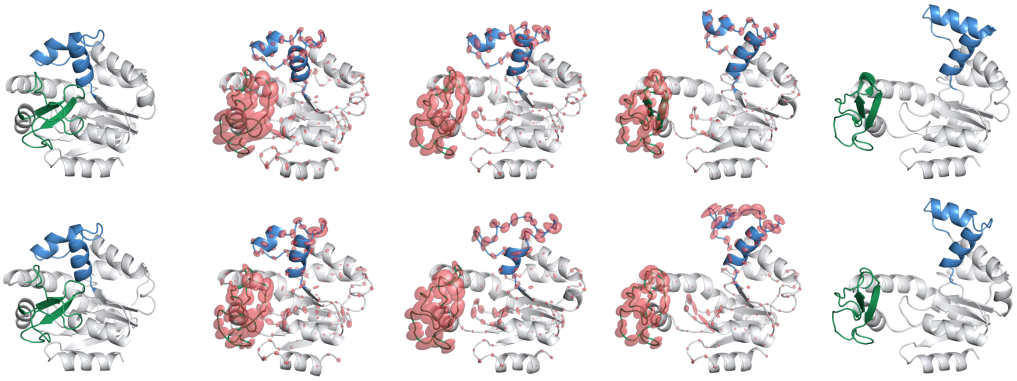


Figure 14: **Learned process geometry in molecular coordinates.** Per-residue mobility along held-out trajectories 111 (top) and 145 (bottom), shown as one-standard-deviation ellipsoids. The **LID** and **NMP** domains exhibit greater predicted mobility than the **CORE**.

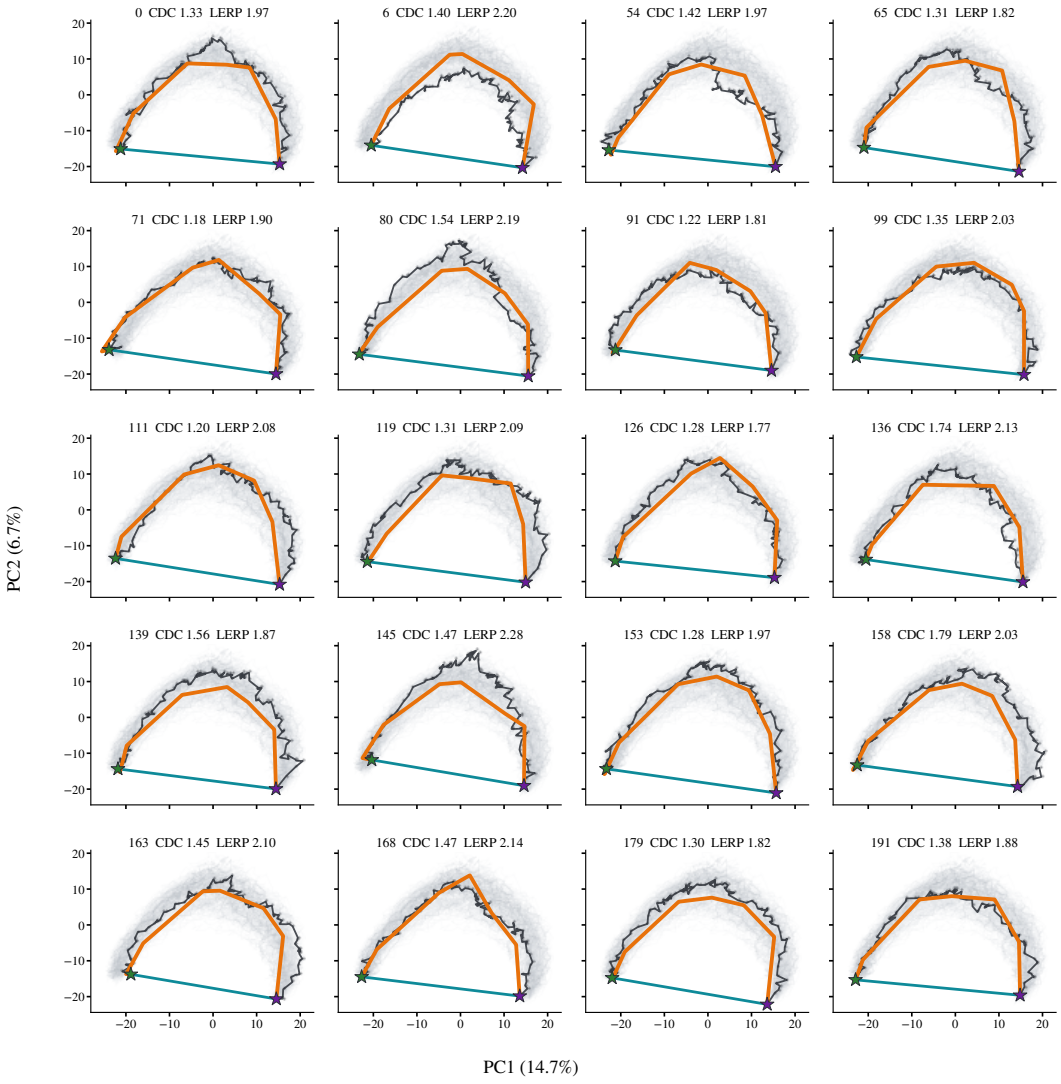



Figure 15: **Endpoint-conditioned paths on held-out MD trajectories.** Each panel compares a held-out trajectory (**grey**), its IGM bridge (**orange**), and latent linear interpolation (LERP; **teal**) between the same closed (★) and open (★) endpoints. Paths are projected onto the first two principal components of the training states. Headings report trajectory IDs and decoded discrete Fréchet distances in Å, using C$\alpha$ Kabsch-RMSD as the ground distance. IGM improves on LERP for all 20 test trajectories, with mean distances of 1.418 and 2.039 Å, respectively. Rank $r = 32$ and lag $\tau = 16$ ps were selected on the validation trajectories.