Identifiable alignment of unpaired representations

Stas Syrota, Albert Lopez I Serrano, Johanne Franck, Søren Hauberg
Technical University of Denmark
{stasy,alise,--,sohau}@dtu.dk

Abstract

Aligning two models without paired data requires a distance that both share, yet existing methods use Euclidean or spherical geometry. This is either an arbitrary choice or imposed by their loss function. We show that, under a precise form of the manifold hypothesis, the pullback of the data metric gives such a distance for any two models of one data-generating process. This holds whether the models are encoders, decoders or both, and whatever their architecture, loss, prior or training data. The alignment map preserves this distance and the distribution of representations, and is identifiable if and only if the data distribution breaks every symmetry of the data manifold. We show that this symmetry breaking holds generically. Identifiability makes the alignment coupling the unique optimum of Gromov–Wasserstein and of its weighted combinations with a linear program. The linear program alone has the same optimum if almost every data point has a distinct distribution of distances. The coupling, and with it the alignment map, is recovered exactly from permuted samples, even without identifiability, and consistently from independent ones. Our method, Neural Gromov–Monge Alignment (NGMA), uses no pairs or labels and outperforms every unpaired and label-informed baseline. It relies on pullback distances to align image models with different priors and single-cell models trained on disjoint cells, where Euclidean distances fail.

1 Introduction

Building multimodal foundation models [Merullo et al., 2023], sharing knowledge across data-privacy walls [Lotfollahi et al., 2022], and auditing proprietary vector-database security [Jha et al., 2025] all rely on learning an *alignment map* between two representation spaces, which sends the representation of each data point in one space to the representation of the same data point in the other. With paired data, this is a regression problem. Without pairs, individual representations are meaningless in absolute terms. What remains in each space are the relative distances between points and their distribution, which unpaired methods use to recover the alignment.

This idea rests on the Platonic representation hypothesis [Huh et al., 2024], which posits that independently trained models converge to the same *similarity structure* on the data, defined as a kernel on pairs of data points computed from their representations. Given two sets of representations, unpaired methods therefore learn a map that matches the kernels, in the hope that it is the alignment.

Recent methods align text encoders whose loss function places the representations on a sphere [Wang and Isola, 2020, Jha et al., 2025, Dar, 2025], where equal kernels make the alignment map a rotation. Their cross-modal counterpart [Schnaus et al., 2026] extends this to vision encoders by imposing a spherical geometry on their representations before alignment. The advances therefore predominantly cover one family of loss functions and two modalities, and whether they extend beyond these is an open problem. The case to settle first is a single modality under arbitrary loss functions, which any account of representational convergence must pass before it can extend across modalities.

The immediate problem is that both signals, the relative distances and the distribution of the representations, depend on the choice of geometry. Intuitively, leveraging them requires a distance tied to the

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

intrinsic geometry of the data, invariant to arbitrary modelling choices. Unfortunately, the field almost universally assumes the Euclidean or spherical geometry for the representations, which the data do not necessarily determine. Standard contrastive losses build the sphere into training [Wang and Isola, 2020]. Balestriero and LeCun [2026] prove that, at their optimum, whitened joint-embedding predictive architectures (JEPAs) trained on the same finite set of data points with the same positive pairs share a metric, but only on those points, not on their generating distribution. In both cases, the loss function sets the geometry, so models trained with different loss functions need not share it. For general loss functions, models can freely reparameterise their representations, which distorts Euclidean distances [Syrota et al., 2025, Hauberg, 2018].

More fundamentally, the literature offers no general account of why trained models should share a similarity structure and what that structure is. Without one, we cannot say how the alignment map is defined, whether a map that matches the kernels is unique, or whether its recovery is well-posed.

In this paper, we show that aligning two models of the same data without pairs is well-posed only when both representations induce the same metric on the data. A formalised manifold hypothesis defines this metric, and treating models as charts of the data manifold defines the alignment map. With this metric, the map is uniquely identifiable if and only if the data distribution breaks the symmetries of the underlying manifold, a condition we prove holds generically. We turn this identifiability into recovery guarantees by casting the problem as transporting one distribution of representations onto the other while preserving pairwise distances. Finally, we instantiate our theory as an architecture-agnostic alignment method for representations of the same dimension. Figure 1 visualises the main contributions of the paper, the relationships between them and how they are presented.

Section 2 states the *metric manifold hypothesis* (H1) and treats each model as a chart of the data manifold (*model coordinates*, H2), so the alignment map is a coordinate change. It preserves latent distances only if both models assign the same distances to data points, as the pullback of the data metric always does. With the distribution of representations, each latent space becomes a *metric measure space* (mm-space) [Gromov, 1999, Sturm, 2006], and the alignment map an isomorphism.

Section 3 shows that A is identifiable, the only isomorphism between the latent mm-spaces, if and only if no isometry of the data manifold other than the identity preserves the data distribution. This *symmetry breaking* (C1) depends on the data-generating process alone and holds generically.

Section 4 recovers A by fitting a map to a coupling of the samples. Under C1, the coupling uniquely solves a non-convex transport problem [Mémoli, 2011]. Under *distinguishable points* (C2), a linear program finds the same optimum. We show that fused GW, which combines the two, keeps the unique optimum under C1 without requiring C2. Recovery is exact when both models encode the same observations and consistent for independent samples.

Section 5 instantiates the theory as Neural Gromov–Monge Alignment (NGMA), which computes the coupling with entropic optimal transport solvers [Peyré et al., 2016, Cuturi, 2013] and fits a map that preserves the pullback metric. **Section 6** tests each claim on the sphere and on image and single-cell models. NGMA outperforms unpaired and label-informed baselines, and the pullback geometry is what makes it work across priors and training cells.

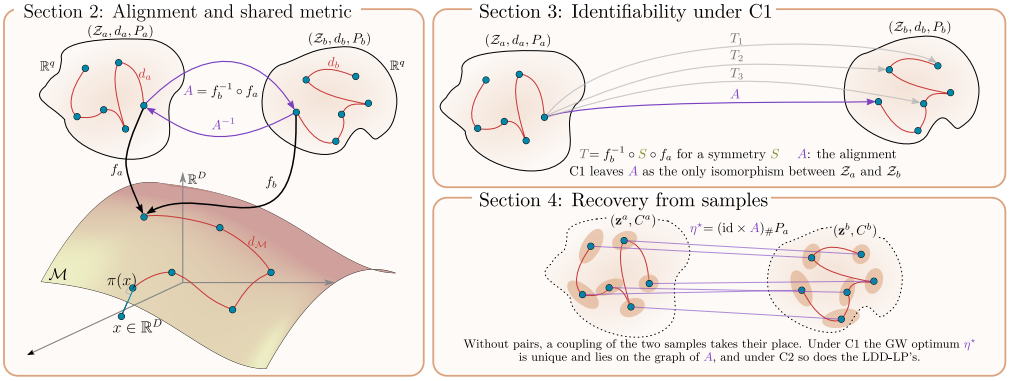


Figure 1: **Outline of Sections 2–4:** the setting, identifiability and recovery from samples.

2 The metric manifold hypothesis defines the alignment and latent geometry

Representation learning is often motivated by the manifold hypothesis [Bengio et al., 2013], which posits that the high-dimensional data concentrates near a lower-dimensional submanifold of the observation space. This idea has proven effective in, e.g., learning causal representations [Schölkopf et al., 2021], biological mechanisms [Lopez et al., 2018], text embeddings [Mikolov et al., 2013], and image generation [Rombach et al., 2022].

Each of these successes uses the learned structure as a space. Clusters are found through distances, latent codes are interpolated along paths, and the representations of two models are compared by how far apart their points lie. All of this implicitly requires the data manifold to be connected, so that any two points are joined, and compact, so that distances stay bounded. It also requires the data to cover the manifold, since a model learns only the part of it that the data reach. Our *metric manifold hypothesis* (H1) states these assumptions explicitly.

Metric manifold hypothesis (H1). Observations $x \sim P_{\mathcal{X}}$ lie in $\mathcal{X} = \mathbb{R}^D$, equipped with a Riemannian metric $g^{\mathcal{X}}$, Euclidean unless stated otherwise. We further assume:

(i) Data concentrate near a submanifold $\mathcal{M} \subset \mathcal{X}$ of dimension $\dim \mathcal{M} < D$ that is smooth, compact and connected, possibly with a boundary [Fefferman et al., 2016, Niyogi et al., 2008].

(ii) The noise-free data follow a distribution $P_{\mathcal{M}}$ with a smooth positive density on $\mathcal{M}$.

(iii) Each observation x has a unique nearest point $\pi(x)$ on $\mathcal{M}$, and it is the noise-free point x came from. This holds when noise moves points perpendicular to $\mathcal{M}$ by less than its reach, the largest distance within which every point has a unique nearest point on $\mathcal{M}$ [Federer, 1959, Aamari et al., 2019].

Distances on $\mathcal{M}$ are measured with the restriction $g^{\mathcal{M}}$ of $g^{\mathcal{X}}$. The resulting shortest-path distance $d_{\mathcal{M}}$ is finite because $\mathcal{M}$ is connected and bounded because it is compact. Compactness also makes the reach positive, so (iii) allows noise of positive size. Projecting the observations onto $\mathcal{M}$ recovers the noise-free data (Remark B.1), so $P_{\mathcal{M}}$ is the distribution of $\pi(x)$ for $x \sim P_{\mathcal{X}}$, the *pushforward* $\pi_{\#}P_{\mathcal{X}}$. A metric space with a probability measure is a *metric measure space*, or mm-space [Gromov, 1999, Sturm, 2006]. We call $(\mathcal{M}, d_{\mathcal{M}}, P_{\mathcal{M}})$ the *data mm-space*.

What it means for a model to represent this structure is also usually left implicit. Using latent distances as distances between data points presupposes that distinct data points have distinct representations and that nearby representations belong to nearby data points. Identifiability theory assumes this explicitly [Hyvärinen and Pajunen, 1999, Khemakhem et al., 2020, Xi and Bloem-Reddy, 2023, Syrota et al., 2025]. Our *model coordinates* hypothesis states it precisely:

Model coordinates (H2). A model provides coordinates on $\mathcal{M}$ through a parameterisation, a diffeomorphism $f\colon \mathcal{Z} \to \mathcal{M}$ from a latent space $\mathcal{Z}$ of the same dimension onto $\mathcal{M}$, that is, a smooth bijection with a smooth inverse. Then $\mathcal{Z}$ inherits the compactness, connectedness and boundary of $\mathcal{M}$. A parameterisation can be built from either an encoder or a decoder (Lemma B.2):

(i) A decoder $\mathrm{dec}\colon \mathbb{R}^q \to \mathcal{X}$, injective with a Jacobian of rank q and image containing $\mathcal{M}$. Then $f = \mathrm{dec}\,|_{\mathcal{Z}}$ is a parameterisation on $\mathcal{Z} := \mathrm{dec}^{-1}(\mathcal{M})$ if $\mathcal{Z}$ is compact.

(ii) An encoder $\mathrm{enc}\colon \mathcal{X} \to \mathbb{R}^q$ with $\mathrm{enc} = \mathrm{enc}\,|_{\mathcal{M}} \circ \pi$, injective on $\mathcal{M}$ with a Jacobian of rank $\dim \mathcal{M}$ there. Then $f = (\mathrm{enc}\,|_{\mathcal{M}})^{-1}$ is a parameterisation on $\mathcal{Z} := \mathrm{enc}(\mathcal{M})$.

Since $\dim \mathcal{M}$ is unknown, decoders usually have $q > \dim \mathcal{M}$, and compactness of $\mathcal{Z}$ keeps neighbouring data points from having distant representations. We write f_a, f_b for the parameterisations of two models and $\mathcal{Z}_a, \mathcal{Z}_b$ for their latent spaces. Two models of the same data are two coordinate systems on one manifold, and aligning them is changing coordinates.

Definition 2.1 (Alignment map)**.** Under H1–H2, the *alignment map* $A\colon \mathcal{Z}_a \to \mathcal{Z}_b$ is:

$$A = f_b^{-1} \circ f_a. \tag{1}$$

A sends each representation in model a to the representation of the same data point in model b.

Although A has a closed form, it cannot be evaluated from unpaired samples and must be inferred from distances and distributions. Both depend on the geometry, and since A only changes coordinates, it preserves a geometry only if both models measure the same distances on the data manifold $\mathcal{M}$.

Proposition 2.2 (A shared data metric)**.** *Under H1–H2, let d_a and d_b be metrics on $\mathcal{Z}_a$ and $\mathcal{Z}_b$. The alignment preserves them, $d_b(A(z), A(z')) = d_a(z, z')$ for all $z, z' \in \mathcal{Z}_a$, if and only if both models*

assign the same distance to every pair of data points,

$$d_a\big(f_a^{-1}(x), f_a^{-1}(y)\big) = d_b\big(f_b^{-1}(x), f_b^{-1}(y)\big) \qquad \textit{for all } x, y \in \mathcal{M}.$$

Euclidean or spherical latent distances satisfy this condition only when A extends to a rigid motion, which is what the methods of Section 1 assume and what the data alone do not guarantee. Balestriero and LeCun [2026] prove that at the optimum, whitened JEPAs trained alike on the same data points share a Euclidean metric on them. This metric satisfies Proposition 2.2, but only models trained with the same loss on the same data points share it.

Latent metric measure spaces. By Proposition 2.2, the metric has to be one that both models induce on $\mathcal{M}$ regardless of their training, and $\mathcal{M}$ already carries a canonical one, $d_{\mathcal{M}}$, from H1. Each model reaches it through its own Jacobian, via the pullback metric $g_a(u, v) := g^{\mathcal{M}}(\mathrm{d}f_a u, \mathrm{d}f_a v)$, written $f_a^* g^{\mathcal{M}}$. The length d_a of the shortest curve in $\mathcal{Z}_a$ under g_a then satisfies $d_a(z, z') = d_{\mathcal{M}}(f_a(z), f_a(z'))$. We pair it with the distribution of the representations, $P_a := (f_a^{-1})_{\#} P_{\mathcal{M}}$, and likewise for b. Two mm-spaces are *isomorphic* if a bijection between them preserves distances and carries one measure onto the other. By construction, f_a is such an isomorphism from the *latent mm-space* $(\mathcal{Z}_a, d_a, P_a)$ onto the data mm-space, and so is f_b. Composing the two gives a key result:

Proposition 2.3. *Under H1–H2, the alignment map is an isomorphism of the latent mm-spaces.*

Unlike the Euclidean metric of whitened JEPAs, the pullback metric is shared by any two models satisfying H2, whatever their architecture, loss function or training points.

3 Identifiability of the alignment map

Proposition 2.3 tells us to search for an isomorphism of the latent mm-spaces, but not whether the search will find A. Without pairs, the samples reveal only distances and distributions, which every such isomorphism, including A, matches. If several exist, the search may return any of them. We call A *identifiable* if it is the only isomorphism. Since both latent mm-spaces inherit their geometry from the data mm-space, identifiability is a property of the data-generating process rather than of the models, and we show that it holds generically.[1]

To illustrate when multiple isomorphisms exist, we build a running example of three data mm-spaces, each the sphere $S^2 \subset \mathbb{R}^3$ with the round metric and a different distribution (Figure 2). The hypersphere is the geometry that contrastive losses [Wang and Isola, 2020] and alignment methods [Schnaus et al., 2026] impose on representations, even when the data manifold itself is not a sphere. The example therefore mirrors such models, but all results hold for any data mm-space satisfying the metric manifold hypothesis (H1).

On each sphere, we take two parameterisations, $f_a = \mathrm{id}$ and $f_b = R^{-1}$ for a fixed rotation R, so model b represents every data point rotated by R and the alignment map is $A = f_b^{-1} \circ f_a = R$. Rotations preserve the round metric, so the latent mm-spaces (S^2, d_{S^2}, P) and $(S^2, d_{S^2}, R_{\#}P)$ are isomorphic. For the uniform distribution, $P = P_{\mathrm{unif}}$ (Figure 2, left), every rotation S other than the identity moves points but leaves the distances between them and their distribution unchanged. The map $f_b^{-1} \circ S \circ f_a = RS$ is then an isomorphism of the latent mm-spaces just like A, and the

[1] A property is generic if an arbitrarily small change can make it hold and no small change can break it. An example is invertibility of square matrices, where a small change of the entries makes a singular matrix invertible and keeps an invertible one invertible. Formally, it holds on an open and dense set.

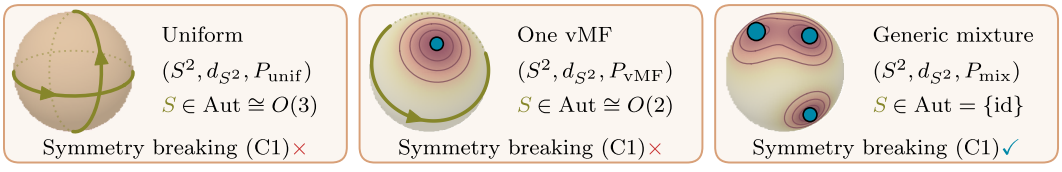


Figure 2: Three data mm-spaces on the round sphere, with the symmetries S that preserve each distribution (olive), the vMF centres (teal, area proportional to weight) and iso-density contours. $O(3)$ consists of all isometries of the sphere, and $O(2)$ of those that fix the vMF centre.

search may return either. Concentrating the data around one point with a von Mises–Fisher (vMF) distribution P_{vMF} [Fisher, 1953] rules out some of these symmetries (middle), and the mixture P_{mix} of three vMF distributions rules out the rest (right), so A is the only isomorphism (Appendix B.4).

A *symmetry* of the data mm-space is a bijection $S\colon \mathcal{M} \to \mathcal{M}$ that preserves distances and $P_{\mathcal{M}}$, and the symmetries form the *automorphism group* $\mathrm{Aut}(\mathcal{M}, d_{\mathcal{M}}, P_{\mathcal{M}})$. Each symmetry S yields an isomorphism $f_b^{-1} \circ S \circ f_a$ that is inseparable from the alignment map without pairs, but differs from A unless S is the identity. The following condition ensures that A is identifiable:

Symmetry breaking (C1). The identity is the only symmetry of the data mm-space.

It is also necessary, since every isomorphism of the latent mm-spaces comes from a symmetry:

Proposition 3.1. *Under H1–H2, the isomorphisms of the latent mm-spaces are exactly the maps $f_b^{-1} \circ S \circ f_a$ with $S \in \mathrm{Aut}(\mathcal{M}, d_{\mathcal{M}}, P_{\mathcal{M}})$, so A is identifiable if and only if C1 holds.*

The condition may seem demanding, since the uniform sphere already violates it. Yet symmetric densities are the exception on any data manifold:

Proposition 3.2. *Fix $\mathcal{M}$ and $g^{\mathcal{M}}$. The densities in $C^\infty_{>0}(\mathcal{M})$ for which the data mm-space satisfies C1 form an open, dense subset in the C^∞ topology.*

A symmetry S is fragile, because the density ρ of $P_{\mathcal{M}}$ must satisfy $\rho \circ S = \rho$ at every point. An arbitrarily small bump at a point that S moves, but not at its image, breaks S. On the sphere, a first vMF bump on P_{unif} breaks every symmetry that moves its centre. The rotations about its axis and the reflections in planes containing it remain, as for P_{vMF} (Figure 2, middle). A second, smaller bump off that axis breaks these rotations, and only the reflection in the plane through both centres remains. A third, even smaller bump off that plane breaks this reflection, which maps the third centre to a point without a bump (right). On any $\mathcal{M}$, well-placed bumps lower the dimension of the continuous symmetries until none remain. Since the isometry group of $\mathcal{M}$ is compact, finitely many symmetries are left, and one more bump breaks each. Densities that satisfy C1 are therefore dense (Appendix B).

Conversely, densities that satisfy C1 form an open set, because a small change cannot create a symmetry. Every isometry of the sphere far from the identity changes the density of P_{mix} somewhere by at least a fixed amount. A small rotation changes it at a rate set by the slope of the density along the rotation, which is nonzero since no rotation preserves P_{mix}. A perturbation whose values and slope stay well below this amount and this rate, such as a slight shift of a centre, creates no symmetry. On any $\mathcal{M}$, the argument uses only that the isometries form a compact Lie group, like $O(3)$.

The pullback metric is what makes identifiability a property of the data-generating process alone. With Euclidean or spherical distances, A preserves them only if A is a rigid motion, which limits alignment to models trained with the same dot-product loss (Section 2). Even then, A is unique only if the models spread their representations without symmetry. Contrastive losses and the LeJEPA regulariser do the opposite. They push the representations towards the uniform distribution on the hypersphere or towards $\mathcal{N}(0, I)$, which every orthogonal map preserves [Wang and Isola, 2020, Balestriero and LeCun, 2025]. Where these losses give two models a shared metric, they do so through this symmetry, which leaves A inseparable from QA for every orthogonal Q.

4 Recovery of the alignment from finite samples

Under symmetry breaking (C1), which holds generically, the latent mm-spaces determine A, so estimating it without pairs is well-posed. We see only independent samples of the two latent mm-spaces, with no one-to-one correspondence between them, so regression is out of reach. Instead, we first estimate a *coupling* and then fit a map to it, akin to neural optimal transport methods [Klein et al., 2024]. We now discuss recovery guarantees for this two-stage procedure.

A coupling is a probability measure $\eta \in \Pi(P_a, P_b)$ on $\mathcal{Z}_a \times \mathcal{Z}_b$ with marginals P_a and P_b. Since the alignment map is an isometry, the coupling induced by A preserves all pairwise distances. We therefore estimate it by solving the *Gromov–Wasserstein* (GW) problem [Mémoli, 2011]:

$$\min_{\eta \in \Pi(P_a, P_b)} J_{\mathrm{GW}}(\eta), \qquad J_{\mathrm{GW}}(\eta) := \iint \left| d_a(z, z') - d_b(w, w') \right|^2 \mathrm{d}\eta(z, w)\, \mathrm{d}\eta(z', w'). \tag{2}$$

It seeks the coupling that distorts distances least, and under C1 it has exactly one solution:

Proposition 4.1. *Under H1–H2, the optimal couplings of Equation 2 are exactly the couplings* $(\mathrm{id} \times T)_{\#} P_a$ *induced by isomorphisms* T *of the latent mm-spaces. Under symmetry breaking (C1), the optimal coupling is unique and equals* $\eta^\star = (\mathrm{id} \times A)_{\#} P_a$.

Since A is continuous and P_a has full support, the coupling $\eta^\star$ is supported on the graph of A, so $\eta^\star$ determines the map. Under C1, both stages therefore have a unique target, $\eta^\star$ for the coupling and A for the map, so the two-stage procedure is well-posed. The proofs are in Appendix B.

4.1 Computing the coupling

The uniqueness in Proposition 4.1, however, does not guarantee that a solver finds the optimum. The GW problem is non-convex [Kravtsova, 2026], so solvers return stationary points that need not be global optima. We therefore turn to Mémoli's third lower bound on GW [Mémoli, 2011], which compares the *local distance distributions* (LDDs) of points instead of their pairwise distances. It is a convex problem, so every stationary point is a global optimum, but it is coarser than GW.

The LDD of a point $z \in \mathcal{Z}_a$ is the distribution function of its distance to a random representation, $\Phi_a(z, r) := P_a\big(\bar{B}_a(z, r)\big)$, where $\bar{B}_a(z, r)$ is the closed ball of radius r around z, and likewise Φ_b on $\mathcal{Z}_b$. Under H1–H2, both latent mm-spaces are isomorphic to the data mm-space, so $\Phi_a(z, \cdot) = \Phi(x, \cdot) = \Phi_b(A(z), \cdot)$ for the data point $x = f_a(z) = f_b(A(z))$, where $\Phi(x, r) := P_{\mathcal{M}}\big(\bar{B}_{\mathcal{M}}(x, r)\big)$ is its LDD in the data mm-space. The alignment map thus preserves LDDs, so we can match them instead of distances by solving

$$\min_{\eta \in \Pi(P_a, P_b)} J_{\mathrm{LDD}}(\eta), \qquad J_{\mathrm{LDD}}(\eta) := \int W_2^2\big(\Phi_a(z, \cdot), \Phi_b(w, \cdot)\big)\, \mathrm{d}\eta(z, w). \tag{3}$$

Equation 3 is a linear program in η, which we call the LDD-LP, and its cost is the squared L^2 distance between the quantile functions of the two LDDs.

Each LDD collapses a point's distances to all others into one distribution, so the LDD-LP tells points apart less finely than GW. If two data points share an LDD, the LDD-LP can exchange them at no cost, so we require almost every point to have a unique LDD:

Distinguishable points (C2). The LDD map $x \mapsto \Phi(x, \cdot)$ is injective $P_{\mathcal{M}}$-almost everywhere, i.e., the collision set $\mathcal{C} := \{x : \Phi(x', \cdot) = \Phi(x, \cdot) \text{ for some } x' \neq x\}$ is $P_{\mathcal{M}}$-null.

C2 implies C1, because a symmetry preserves every LDD, but not conversely, since almost every point can share its LDD with another without a symmetry (Appendix C). If the data mm-space has distinguishable points, we can use the LDD-LP to find the coupling $\eta^\star$:

Proposition 4.2. *Under H1–H2 and C2,* $\eta^\star$ *is the unique optimal coupling of the LDD-LP.*

The two problems trade the fidelity of GW against the convexity of the LDD-LP, and neither has both. We propose to combine them as in fused GW [Vayer et al., 2019], with the LDD-LP in place of the feature term and a weight $\beta \in [0, 1]$ on it:

$$J_\beta(\eta) := (1 - \beta)\, J_{\mathrm{GW}}(\eta) + \beta\, J_{\mathrm{LDD}}(\eta). \tag{4}$$

Both terms are nonnegative and vanish at $\eta^\star$, so under C1, $\eta^\star$ is the unique minimiser of J_β for $\beta < 1$ (Proposition B.17), and under C2 also for $\beta = 1$. Therefore, a solver can lower β from 1 to 0 without moving the minimiser, but J_β is non-convex for $\beta < 1$, and we do not prove that the path ends at $\eta^\star$.

Since LDDs are estimated from samples and carry sampling error, they must be not only distinct but well separated. C2 itself cannot be checked from samples, but the smallest distance between the LDDs of two data points at least t apart can be estimated:

Definition 4.3. For $0 < t \leq \mathrm{diam}(\mathcal{M})$, the *margin* of the data mm-space at separation t is

$$m(t) := \inf\big\{\, \|\Phi(x, \cdot) - \Phi(x', \cdot)\|_\infty \;:\; x, x' \in \mathcal{M},\ d_{\mathcal{M}}(x, x') \geq t \,\big\}. \tag{5}$$

From samples, the DKW inequality [Dvoretzky et al., 1956, Massart, 1990] gives a level-δ margin test of whether two query points at separation t share an LDD. A rejection says nothing about the pairs the test does not query, so it can occur where $m(t) = 0$. In low dimension, however, the queries can cover $\mathcal{M}$, and the estimate then bounds $m(t)$ from below if the LDDs are Lipschitz (Appendix C.3).

4.2 Finite-sample guarantees

In practice we see only samples, $\mathbf{z}^a = (\mathrm{z}_1^a, \dots, \mathrm{z}_n^a)$ in $\mathcal{Z}_a$ and $\mathbf{z}^b$ in $\mathcal{Z}_b$, with distance matrices $C_{ij}^a = d_a(\mathrm{z}_i^a, \mathrm{z}_j^a)$ and C^b. The samples define the empirical distributions $\hat{P}_a$ and $\hat{P}_b$, and with the distance matrices, they carry the two signals that remain without pairs and determine the empirical LDDs. We minimise J_β for some $\beta \in [0,1]$ over the couplings of $\hat{P}_a$ and $\hat{P}_b$, which gives the empirical version of the problem.

In the *independent regime*, the two samples are drawn independently from P_a and P_b, so almost surely no point of one is the image of a point of the other. In the *permutation regime*, both models encode the same observations in unknown order, so that $\mathrm{z}_{\sigma(i)}^b = A(\mathrm{z}_i^a)$ for a permutation σ.

The permutation regime is the easier one, because A is an isometry and the two distance matrices therefore agree up to the unknown order, $C_{\sigma(i)\sigma(k)}^b = C_{ik}^a$. Almost surely the pairwise distances of the sample are distinct, so only σ reorders one matrix into the other or pairs points with equal empirical LDDs (Lemma B.11). The empirical problem therefore recovers σ exactly for every β:

Proposition 4.4. *In the permutation regime, under H1–H2 and with $n \geq 3$, the empirical J_β almost surely has the unique optimal coupling $\frac{1}{n}\sum_i \delta_{(\mathrm{z}_i^a, \mathrm{z}_{\sigma(i)}^b)}$ for every $\beta \in [0,1]$.*

The proposition needs neither C1 nor C2, so it holds even for the uniform distribution P_{unif} on the sphere (Figure 2), where no method identifies A from independent samples. With σ recovered, fitting A is supervised regression, and $\dim \mathcal{M} + 1$ pairs fix the alignment among isometries if they are in general position and close enough that the shortest paths between them are unique (Corollary A.7).

In the independent regime, no coupling of the two samples is exact at finite n, and only consistency remains (Proposition B.14). Global minimisers of the empirical J_β converge to $\eta^\star$ under C1 for $\beta < 1$, and under C2 also for $\beta = 1$, by compactness and the uniqueness of the population minimiser (Propositions 4.1, 4.2 and B.17). In each case, a map fitted to the coupling over a uniformly compact class containing A converges to A. Consistency leaves two questions open. First, it covers only global minimisers, while solvers return stationary points that need not be global. Second, it gives no rate of convergence.

5 Neural Gromov–Monge Alignment (NGMA)

Our Neural Gromov–Monge Alignment (NGMA) method instantiates the two-stage recovery procedure with choices informed by Section 4. Its input is the samples $\mathbf{z}^a$, $\mathbf{z}^b$, their distance matrices C^a, C^b, and the pullback metric tensors G_a, G_b at the sample points.

NGMA computes the coupling $\hat{\eta}$ with entropic GW [Peyré et al., 2016] applied to J_β in four stages, $\beta = 1, 0.8, 0.5$ and 0, that lower the entropy weight together with β, each starting from the coupling of the previous one (Appendix D). The first stage is the convex LDD-LP at $\beta = 1$ and the last is entropic GW at $\beta = 0$, so NGMA returns a stationary point of entropic GW, reached from the solution of the convex problem rather than from the independent coupling.

Given $\hat{\eta}$, NGMA estimates A, a smooth isometry that carries P_a onto P_b (Proposition 2.3). Such a map solves the non-convex Gromov–Monge problem [Mémoli and Needham, 2022], which NGMA avoids by fitting a smooth map to $\hat{\eta}$, the coupling that Section 4 recovers.

A smooth isometry is also a bijection. Architectures with a closed-form inverse would build this in, but they restrict the class of maps, so NGMA fits two plain networks, ϕ for A and ψ for A^{-1}. Each is an affine map plus a residual network, with the affine part initialised at the Procrustes solution $\min_{Q,v} \sum_{i,j} \hat{\eta}_{ij} \|Q\mathrm{z}_i^a + v - \mathrm{z}_j^b\|^2$ [Schönemann, 1966], or its inverse for ψ, and the residual at zero. Both are trained on $\mathcal{L}(\phi,\psi) = \ell_{ab}(\phi,\psi) + \ell_{ba}(\psi,\phi)$, where

$$\ell_{ab}(\phi,\psi) = \frac{1}{n_a}\sum_i \left[\frac{\|J_\phi^\top G_b\, J_\phi - G_a\|_F^2}{\|G_a\|_F^2} + \lambda_{\mathrm{fit}} \|\phi(\mathrm{z}_i^a) - \tau_i^a\|^2 + \lambda_{\mathrm{cyc}} \|\psi(\phi(\mathrm{z}_i^a)) - \mathrm{z}_i^a\|^2 \right], \quad (6)$$

J_ϕ and G_a are evaluated at z_i^a, G_b at the point of $\mathbf{z}^b$ nearest to $\phi(\mathrm{z}_i^a)$, τ_i^a is the z_j^b with the largest $\hat{\eta}_{ij}$, and ℓ_{ba} exchanges the roles of a and b. Weights are set at $\lambda_{\mathrm{fit}} = 0.3$ and $\lambda_{\mathrm{cyc}} = 0.1$.

The first term penalises deviations from $A^* g_b = g_a$, the last pushes ψ towards ϕ^{-1}, and among maps that nearly satisfy both, the fit term favours the one that follows the coupling, which at $\eta^\star$ is A.

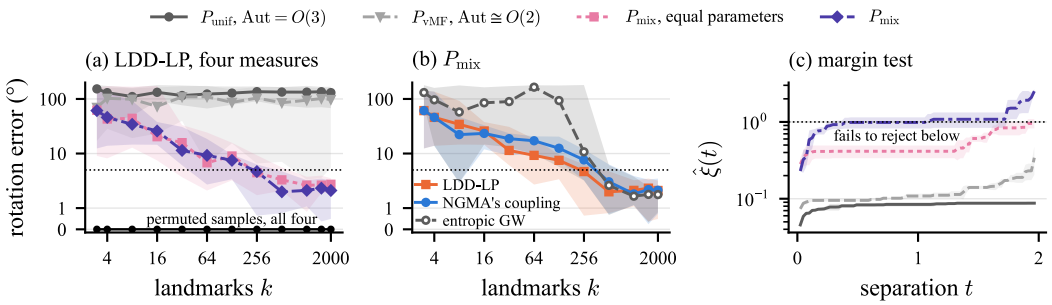


Figure 3: Data on the sphere with $A = R$. **(a)** Rotation error of the LDD-LP on independent samples against the number of landmarks k (median and range over 10 seeds; at $k = 2000$ every point is a landmark), and zero on permuted samples (black). **(b)** The LDD-LP, NGMA's coupling and entropic GW alone, annealed as in NGMA, on P_{mix}. **(c)** The margin test: the estimated margin over its DKW accuracy, $\hat{\xi}(t)$, mean $\pm$ s.d. over 3 seeds, which rejects equal LDDs where $\hat{\xi}(t) > 1$.

We hold out n_0 random points of $\mathbf{z}^a$ and every point of $\mathbf{z}^b$ whose largest entry of $\hat{\eta}$ falls on one of them, train on the remaining n_a and n_b points, and keep the checkpoint at which the most held-out points map to a mutual nearest neighbour in the other sample. The coupling's targets are wrong for many individual points, so, as in unsupervised word translation [Conneau et al., 2018, Artetxe et al., 2018], we refit both maps for up to three rounds on the pairs they agree on, the mutual nearest neighbours under $D(i,j) = \omega_a\|\phi(\mathrm{z}_i^a) - \mathrm{z}_j^b\| + \omega_b\|\psi(\mathrm{z}_j^b) - \mathrm{z}_i^a\|$. Both distances are divided by their medians, and each weight is proportional to the squared fraction of mutual nearest neighbours its map finds. We keep the round with the most held-out mutual nearest neighbours under D and rank test candidates by D (Appendix D). Code is available at `https://github.com/mustass/identifiable-alignment`.

6 Experiments

First, we test the claims of Sections 2–5 on the sphere, where we choose A, and the geometry is in closed form. Then we train VAEs [Kingma and Welling, 2014] on three image datasets (MNIST [LeCun et al., 1998], FashionMNIST [Xiao et al., 2017] and CelebA [Liu et al., 2015]) with a Gaussian prior or a VampPrior [Tomczak and Welling, 2018], and scVI models [Lopez et al., 2018] on Tabula Sapiens lung cells [The Tabula Sapiens Consortium, 2022].

On real data, we split each dataset into three disjoint subsets: one to train the models, one to fit the alignment and one to test it. All image models are trained on the same training set, across priors and seeds, while the lung models are trained on disjoint sets of cells. We draw the two samples from the alignment subset, in regimes ranging from disjoint to fully shared (Section 4.2), and repeat this with different seeds. NGMA and the baselines are fitted to each draw and scored on the test subset by instance retrieval (R@1), the fraction of test points whose own representation in the other model is ranked first. For all models, we measure latent distances with the pullback metric of an ensemble of eight decoders [Syrota et al., 2024], pulling back the Euclidean metric of pixel space for the images and the Fisher metric of scVI's negative-binomial likelihood for the cells. Conceptually, the ensemble keeps the geometry on the compact latent space $\mathcal{Z}$ of H2, since its metric grows where the decoders disagree, which is away from the data [Hauberg, 2018, Syrota et al., 2024]. Instead of computing geodesics, we use the lengths of straight segments under this metric, which bound the geodesic distances from above, need no optimisation per pair and lose little accuracy (Appendix E.4). On the sphere, the two samples are drawn from the measure itself, and we score the fitted rotation against the true R. Each paragraph below states one finding, with the main results in Table 1 and Figures 3 and 4.

Symmetry breaking (C1) decides identifiability. On each sphere of Figure 2, model a represents the data as they are and model b rotates them by a fixed R, so $A = R$. We add a mixture with equal weights and concentrations, which satisfies C1 but whose LDDs collide along curves, so its margin is zero (Appendix C). The NGMA coupling, the LDD-LP and GW are solved on k landmarks, chosen so that their LDDs differ as much as possible (Appendix E.1). From independent samples, the LDD-LP recovers R to within $5°$ for both mixtures, while for P_{unif} and P_{vMF} the error stays

Table 1: R@1 among the 10,000 test points (19,962 on CelebA) of VAEs aligned on disjoint samples ($n = 3000$ per side, 7500 on CelebA), as mean $\pm$ s.d. over model pairings and draws. The best unpaired baseline is Rosetta, or mini-vec2vec where marked †. The best label-informed baseline is RelGeo, or RelReps where marked ‡. The paired ceilings fit NGMA's map and a rigid map on the true pairs. On CelebA, parentheses count draws reaching R@1 of 0.3.

			Best baseline		Paired ceiling	
	Prior	NGMA	unpaired	label-informed	NGMA	Rigid map
MNIST	Gaussian	**0.967**$\pm$0.021	0.941$\pm$0.028	0.525$\pm$0.069‡	0.998$\pm$0.001	0.955$\pm$0.021
	VampPrior	**0.989**$\pm$0.012	0.930$\pm$0.012	0.552$\pm$0.235	1.000$\pm$0.000	0.936$\pm$0.008
	cross-prior	**0.955**$\pm$0.008	0.571$\pm$0.021	0.319$\pm$0.152	0.999$\pm$0.000	0.512$\pm$0.021
Fashion-MNIST	Gaussian	**0.719**$\pm$0.112	0.710$\pm$0.040	0.159$\pm$0.134	0.997$\pm$0.001	0.782$\pm$0.032
	VampPrior	**0.979**$\pm$0.027	0.656$\pm$0.045	0.355$\pm$0.131	1.000$\pm$0.000	0.747$\pm$0.029
	cross-prior	**0.871**$\pm$0.049	0.065$\pm$0.019	0.098$\pm$0.033	0.998$\pm$0.001	0.071$\pm$0.004
CelebA	Gaussian	**0.804**$\pm$0.375 (12/15)	0.567$\pm$0.453 (8/15)	0.445$\pm$0.033‡	0.995$\pm$0.000	0.993$\pm$0.000
	VampPrior	**0.818**$\pm$0.372 (8/10)	0.616$\pm$0.415† (7/10)	0.387$\pm$0.139	0.996$\pm$0.000	0.995$\pm$0.000
	cross-prior	**0.818**$\pm$0.372 (8/10)	0.757$\pm$0.372 (8/10)	0.220$\pm$0.120	0.996$\pm$0.000	0.990$\pm$0.000

large however many landmarks we use (Figure 3a). Symmetry breaking thus decides recovery, as Proposition 3.1 states, even for the mixture whose margin is zero. The margin test never rejects for the two symmetric measures (Figure 3c, Appendix C.3), but has limited power.

With the pullback metric, models are coordinate charts (H2). On MNIST and FashionMNIST, models with different priors differ by 10 to 25 nats in likelihood (Table 5). Yet their pullback distances correlate above 0.97, while their Euclidean distances correlate at only 0.71 and 0.61 (Table 6). Here, the alignment is highly nonrigid and every method that fits a rotation falls behind NGMA (Table 1). The lung models are not rigidly aligned either, since a rigid map fitted on the true pairs stays far below a map fitted on them (Figure 4).

The permutation regime is easy with the shared metric (Proposition 4.4). With permuted samples, LDD-LP recovers R exactly for every measure, including the uniform one, from 3 landmarks on, as the proposition predicts. When both models encode the same images, NGMA reaches R@1 of 0.990–0.999 in every MNIST and FashionMNIST setting. In contrast, with a GW coupling on Euclidean distances, it reaches only 0.353 and 0.679 on the cross-prior pairs (Appendix E.3).

NGMA leads in every setting, from disjoint to fully shared samples. Our baselines without pairs or labels are Rosetta [Schnaus et al., 2026], mini-vec2vec [Dar, 2025], GENOT-Q [Klein et al., 2024] and LFM [Fumero et al., 2024], and label-informed RelReps [Moschella et al., 2023] and RelGeo [Yu et al., 2025]. In the hardest regime, the two samples share no image, so no coupling is exact and only consistency remains (Section 4.2). Yet NGMA leads every baseline in all nine settings of Table 1, and keeps the edge when samples have images in common (Appendix E.3, Table 7).

Identifiability is a property of the data-generating process, not the training data. Nothing in H1–H2 requires the two models to share training points, so we align lung models trained on disjoint sets of cells, which share only the tissue (Figure 4, Appendix E.6). Without pairs, our method retrieves the true partner of up to 38% of ten thousand held-out cells, where chance is one in ten thousand, and the right cell type for nine in ten. NGMA is above every baseline on every sample, RelGeo with cell-type labels included. With Euclidean distances, it fails on most draws. With the pullback metric of the scVI decoder alone, instead of the ensemble, it is lower on 17 of 18 draws.

Near-symmetries obstruct alignment and fused GW (Equation 4) helps resolve them. On CelebA, a left-right flip nearly preserves the distribution of faces, so C1 nearly fails. On every failed draw, with either coupling, the maps send faces to their mirror images (mirror R@1 of 0.87–0.99, true R@1 at most 0.11). On every successful draw, they send faces to themselves (Appendix E.5). The mirror alignment preserves distances and nearly preserves the distribution of representations, so without pairs it is almost indistinguishable from A. Of the ten draws that fail with GW alone, NGMA's coupling rescues three and loses none, so it succeeds on 28 of 35 (Figure 5). On P_{mix} on the sphere, GW alone leaves 4 to 8 of 10 seeds more than 90° from R at every k up to 256. NGMA's coupling has no such seed from $k = 8$, thanks to its LDD-LP term, although the LDD-LP alone has the smaller median error for k between 32 and 256 (Figure 3b).

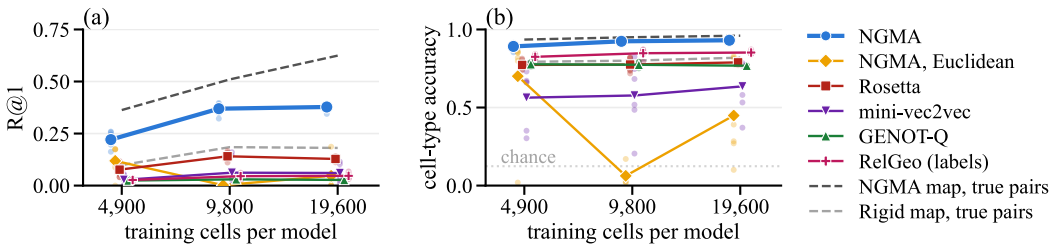


Figure 4: scVI models of lung cells trained on different cells, aligned on disjoint samples of 3000 cells per side, against the number of training cells per model. **(a)** R@1 among 10,000 held-out cells (chance 10^{-4}). **(b)** Cell-type accuracy of the nearest cell. Lines are means over two model pairings and three draws, faint markers are single draws, and dashed lines are maps fitted on the true pairs.

7 Related work

Model stitching and relative representations align two models through pairs, anchors or class labels [Csiszárik et al., 2021, Moschella et al., 2023, Maiorca et al., 2023, Lähner and Moeller, 2024], and Yu et al. [2025] build them from pullback geodesic distances to anchors. Without them, unsupervised word translation and its successors for embeddings fit an orthogonal map, which assumes Euclidean or cosine geometry [Conneau et al., 2018, Artetxe et al., 2018, Grave et al., 2019, Schnaus et al., 2026, Dar, 2025]. vec2vec fits a map with adversarial and cycle losses and a loss that preserves pairwise inner products [Jha et al., 2025]. These methods do not ask whether the recovered map is unique, and the recent ones rest on the Platonic representation hypothesis [Huh et al., 2024]. We formalise the manifold hypothesis, which gives both models a shared metric, the pullback metric, and makes the alignment map a measure-preserving isometry, identifiable under symmetry breaking (C1).

Gromov–Wasserstein matching aligns two samples without correspondence [Mémoli, 2011, Alvarez-Melis and Jaakkola, 2018, Demetci et al., 2022, Schnaus et al., 2025]. Klein et al. [2024] handle the non-uniqueness of GW couplings empirically, and we state when the optimum is unique (Proposition 4.1). The LDD-LP is Mémoli's third lower bound on GW with the squared W_2 cost, whose entropic version Sato et al. [2020] solve with Sinkhorn. Hur and Khoo [2025] match points with its W_1 cost, whose unique optimum is the true matching if no two points share a distance profile, a finite form of C2. Piening and Beinert [2026] use the bound in place of GW in fused GW [Vayer et al., 2019], and Raviv et al. [2012] add a distance-histogram cost to GW over bijections.

Identifiability theory asks when the latent variables of one model are determined up to a set of indeterminacies [Hyvärinen and Pajunen, 1999, Khemakhem et al., 2020, Xi and Bloem-Reddy, 2023]. Arvanitidis et al. [2018] measure latent distances with the pullback metric of the decoder. Syrota et al. [2025] show that these distances are identifiable even when the coordinates are not. We study the map between two models, which needs neither model to be identifiable. Ebin [1970] shows that a generic metric has no isometries other than the identity. We fix the metric and show the same genericity over measures (Proposition 3.2).

8 Discussion

We show that any two models of one data-generating process are related by a single alignment map, an isometry of their pullback geometries that carries one distribution of representations onto the other. This holds whether a model is built from an encoder, a decoder or both, and whatever its architecture, loss, prior or training data. The map is identifiable exactly when the data break all symmetries, which generic data do, and two samples of representations recover it exactly if both models encode the same observations and consistently if they do not. NGMA uses this to align image and single-cell models without pairs, ahead of every baseline without true pairs, across priors and disjoint training data.

Limitations. The theory covers any models of one data-generating process, but our recovery guarantees hold only for global optima, which NGMA's solver may not reach. NGMA's maps act on whole latent spaces of equal dimension, as in all our experiments. The latent dimension is a modelling choice, since that of the data manifold is unknown, and two models can choose differently.

Acknowledgements. This work was supported by a research grant (42062) from VILLUM FONDEN. This project received funding from the European Research Council (ERC) under the European Union's Horizon research and innovation programme (grant agreement 101125993). The work was partly funded by the Novo Nordisk Foundation through the Center for Basic Machine Learning Research in Life Science (NNF20OC0062606). We acknowledge [DTU Computing Center, 2024] for awarding this project access to the LUMI supercomputer, owned by the EuroHPC Joint Undertaking, hosted by CSC (Finland) and the LUMI consortium.

References

Eddie Aamari, Jisu Kim, Frédéric Chazal, Bertrand Michel, Alessandro Rinaldo, and Larry Wasserman. Estimating the reach of a manifold. *Electronic Journal of Statistics*, 13(1):1359 – 1399, 2019. doi: 10.1214/19-EJS1551. URL `https://doi.org/10.1214/19-EJS1551`.

David Alvarez-Melis and Tommi Jaakkola. Gromov-Wasserstein alignment of word embedding spaces. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1881–1890, 2018. doi: 10.18653/v1/d18-1214.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 789–798, 2018. doi: 10.18653/v1/p18-1073.

Georgios Arvanitidis, Lars Kai Hansen, and Søren Hauberg. Latent space oddity: on the curvature of deep generative models. In *International Conference on Learning Representations*, 2018. doi: 10.48550/arxiv.1710.11379.

Randall Balestriero and Yann LeCun. LeJEPA: Provable and scalable self-supervised learning without the heuristics. arXiv preprint arXiv:2511.08544, 2025.

Randall Balestriero and Yann LeCun. Proving and improving the Platonic representation hypothesis with JEPAs. arXiv preprint arXiv:2610.39302, 2026.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35:1798–1828, 2013. doi: 10.1109/tpami.2013.50.

Patrick Billingsley. *Convergence of Probability Measures*. Wiley Series in Probability and Statistics. John Wiley & Sons, 2 edition, 1999. ISBN 9780471197454. doi: 10.1002/9780470316962.

Isaac Chavel. *Eigenvalues in Riemannian Geometry*, volume 115 of *Pure and Applied Mathematics*. Academic Press, 1984. ISBN 9780121706401.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. Word translation without parallel data. In *International Conference on Learning Representations*, 2018. doi: 10.48550/arxiv.1710.04087.

Adrián Csiszárik, Péter Kőrösi-Szabó, Ákos K. Matszangosz, Gergely Papp, and Dániel Varga. Similarity and matching of neural network representations. In *Advances in Neural Information Processing Systems*, volume 34, pages 5656–5668, 2021. doi: 10.48550/arxiv.2110.14633.

Marco Cuturi. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26, pages 2292–2300, 2013. doi: 10.48550/arxiv.1306.0895.

Guy Dar. mini-vec2vec: Scaling universal geometry alignment with linear transformations. arXiv preprint arXiv:2510.02348, 2025. URL `https://arxiv.org/abs/2510.02348`.

Pinar Demetci, Rebecca Santorella, Björn Sandstede, William Stafford Noble, and Ritambhara Singh. SCOT: Single-cell multi-omics alignment with optimal transport. *J. Comput. Biol.*, 29:3–18, 2022. doi: 10.1089/cmb.2021.0446.

Manfredo P. do Carmo. *Riemannian Geometry*. Mathematics: Theory and Applications. Birkhäuser, 1992. ISBN 9780817634902.

DTU Computing Center. DTU Computing Center resources, 2024. URL https://doi.org/10.48714/DTU.HPC.0001.

A. Dvoretzky, J. Kiefer, and J. Wolfowitz. Asymptotic Minimax Character of the Sample Distribution Function and of the Classical Multinomial Estimator. *The Annals of Mathematical Statistics*, 27:642–669, 1956. doi: 10.1214/aoms/1177728174. URL https://doi.org/10.1214/aoms/1177728174.

David G. Ebin. The manifold of Riemannian metrics. In Shiing-Shen Chern and Stephen Smale, editors, *Global Analysis*, volume 15 of *Proceedings of Symposia in Pure Mathematics*, pages 11–40. American Mathematical Society, 1970. ISBN 9780821814154. doi: 10.1090/pspum/015/0267604.

Lawrence C. Evans and Ronald F. Gariepy. *Measure Theory and Fine Properties of Functions*. Textbooks in Mathematics. CRC Press, revised edition, 2015. ISBN 9781482242386.

Herbert Federer. Curvature measures. *Transactions of the American Mathematical Society*, 93:418–491, 1959. doi: 10.1090/s0002-9947-1959-0110078-1.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016. doi: 10.1090/jams/852. URL https://doi.org/10.1090/jams/852.

R. A. Fisher. Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, 217:295–305, 1953. ISSN 1364-5021. doi: 10.1098/rspa.1953.0064. URL https://doi.org/10.1098/rspa.1953.0064.

Rémi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurélie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Léo Gautheron, Nathalie T. H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python optimal transport. *J. Mach. Learn. Res.*, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in Wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162:707–738, 2015. doi: 10.1007/s00440-014-0583-7.

Marco Fumero, Marco Pegoraro, Valentino Maiorca, Francesco Locatello, and Emanuele Rodolà. Latent functional maps: a spectral framework for representation alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pages 66178–66203, 2024. doi: 10.52202/079017-2115.

Edouard Grave, Armand Joulin, and Quentin Berthet. Unsupervised alignment of embeddings with Wasserstein Procrustes. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 1880–1890. PMLR, 2019.

Alfred Gray. The volume of a small geodesic ball of a Riemannian manifold. *Michigan Math. J.*, 20:329–344, 1974. doi: 10.1307/mmj/1029001150.

Misha Gromov. *Metric Structures for Riemannian and Non-Riemannian Spaces*, volume 152 of *Progress in Mathematics*. Birkhäuser, 1999. ISBN 9780817638986.

Brian C. Hall. *Lie Groups, Lie Algebras, and Representations: An Elementary Introduction*, volume 222 of *Graduate Texts in Mathematics*. Springer International Publishing, 2 edition, 2015. ISBN 9783319134666. doi: 10.1007/978-3-319-13467-3.

Søren Hauberg. Only Bayes should learn a manifold (on the estimation of differential geometric structure from data). arXiv preprint arXiv:1806.04994, 2018.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (GELUs). *arXiv preprint arXiv:1606.08415*, 2016. doi: 10.48550/arxiv.1606.08415.

Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. Position: The Platonic representation hypothesis. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 20617–20642. PMLR, 2024. URL https://proceedings.mlr.press/v235/huh24a.html.

YoonHaeng Hur and Yuehaw Khoo. Robust point matching with distance profiles. *Journal of Machine Learning Research*, 26(205):1–38, 2025. URL http://jmlr.org/papers/v26/24-2224.html.

Aapo Hyvärinen and Petteri Pajunen. Nonlinear independent component analysis: Existence and uniqueness results. *Neural Networks*, 12(3):429–439, 1999. doi: 10.1016/S0893-6080(98)00140-3.

Rishi Jha, Collin Zhang, Vitaly Shmatikov, and John X. Morris. Harnessing the universal geometry of embeddings. In *Advances in Neural Information Processing Systems*, volume 38, pages 51335–51359, 2025. doi: 10.52202/085713-1530. URL https://openreview.net/forum?id=jiCLUPq5xv.

Ilyes Khemakhem, Diederik Kingma, Ricardo Monti, and Aapo Hyvärinen. Variational autoencoders and nonlinear ICA: A unifying framework. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pages 2207–2217. PMLR, 2020.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

Diederik P. Kingma and Max Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014. doi: 10.48550/arxiv.1312.6114.

Dominik Klein, Théo Uscidda, Fabian Theis, and Marco Cuturi. GENOT: Entropic (Gromov) Wasserstein flow matching with applications to single-cell genomics. In *Advances in Neural Information Processing Systems 37*, pages 103897–103944, 2024. doi: 10.52202/079017-3301.

Shoshichi Kobayashi and Katsumi Nomizu. *Foundations of Differential Geometry, Volume I*. Wiley Classics Library. John Wiley & Sons, 1996. ISBN 9780471157335.

Steven G. Krantz and Harold R. Parks. *A Primer of Real Analytic Functions*. Birkhäuser Advanced Texts Basler Lehrbücher. Birkhäuser, 2 edition, 2002. ISBN 9780817642648. doi: 10.1007/978-0-8176-8134-0. URL https://doi.org/10.1007/978-0-8176-8134-0.

Natalia Kravtsova. Note on computational complexity of the Gromov-Wasserstein distance. *arXiv preprint arXiv:2408.06525*, 2026. doi: 10.48550/arxiv.2408.06525. v4. Earlier versions were titled "The NP-hardness of the Gromov-Wasserstein distance"; v4 proves non-convexity and gives no NP-hardness proof.

Zorah Lähner and Michael Moeller. On the direct alignment of latent spaces. In *Proceedings of UniReps: the First Workshop on Unifying Representations in Neural Models*, volume 243 of *Proceedings of Machine Learning Research*, pages 158–169. PMLR, 2024.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998. doi: 10.1109/5.726791.

John M. Lee. *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. Springer, 2 edition, 2012. ISBN 9781441999818.

John M. Lee. *Introduction to Riemannian Manifolds*, volume 176 of *Graduate Texts in Mathematics*. Springer, 2 edition, 2018. ISBN 9783319917542. doi: 10.1007/978-3-319-91755-9.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pages 3730–3738, 2015. doi: 10.1109/ICCV.2015.425.

Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 15:1053–1058, 2018. doi: 10.1038/s41592-018-0229-2. URL https://doi.org/10.1038/s41592-018-0229-2.

Mohammad Lotfollahi, Mohsen Naghipourfar, Malte D. Luecken, Matin Khajavi, Maren Büttner, Marco Wagenstetter, Žiga Avsec, Adam Gayoso, Nir Yosef, Marta Interlandi, Sergei Rybakov, Alexander V. Misharin, and Fabian J. Theis. Mapping single-cell data to reference atlases by transfer learning. *Nature Biotechnology*, 40(1):121–130, 2022. doi: 10.1038/s41587-021-01001-7.

Valentino Maiorca, Luca Moschella, Antonio Norelli, Marco Fumero, Francesco Locatello, and Emanuele Rodolà. Latent space translation via semantic alignment. In *Advances in Neural Information Processing Systems*, volume 36, pages 55394–55414, 2023. doi: 10.52202/075280-2418.

P. Massart. The Tight Constant in the Dvoretzky-Kiefer-Wolfowitz Inequality. *Ann. Probab.*, 18: 1269–1283, 1990. doi: 10.1214/aop/1176990746. URL `https://doi.org/10.1214/aop/1176990746`.

Facundo Mémoli. Gromov–Wasserstein distances and the metric approach to object matching. *Found. Comput. Math.*, 11:417–487, 2011. doi: 10.1007/s10208-011-9093-5. URL `https://doi.org/10.1007/s10208-011-9093-5`.

Facundo Mémoli and Tom Needham. Distance distributions and inverse problems for metric measure spaces. *Stud. Appl. Math.*, 149:943–1001, 2022. doi: 10.1111/sapm.12526.

Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space. In *International Conference on Learning Representations*, 2023.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations (ICLR), Workshop Track*, 2013. URL `https://arxiv.org/abs/1301.3781`.

B. S. Mityagin. The zero set of a real analytic function. *Mathematical Notes*, 107:529–530, 2020. doi: 10.1134/s0001434620030189. URL `https://doi.org/10.1134/S0001434620030189`.

Luca Moschella, Valentino Maiorca, Marco Fumero, Antonio Norelli, Francesco Locatello, and Emanuele Rodolà. Relative representations enable zero-shot latent space communication. In *International Conference on Learning Representations*, 2023. doi: 10.48550/arxiv.2209.15430. URL `https://openreview.net/forum?id=SrC-nwieGJ`.

S. B. Myers and N. E. Steenrod. The group of isometries of a Riemannian manifold. *Annals of Mathematics*, 40:400–416, 1939. doi: 10.2307/1968928.

Partha Niyogi, Stephen Smale, and Shmuel Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete & Computational Geometry*, 39(1-3):419–441, 2008. ISSN 1432-0444. doi: 10.1007/s00454-008-9053-2. URL `https://doi.org/10.1007/s00454-008-9053-2`.

Richard S. Palais. On the differentiability of isometries. *Proc. Amer. Math. Soc.*, 8:805–807, 1957. doi: 10.1090/s0002-9939-1957-0088000-x.

Peter Petersen. *Riemannian Geometry*, volume 171 of *Graduate Texts in Mathematics*. Springer International Publishing, 3rd edition, 2016. ISBN 9783319266527. doi: 10.1007/978-3-319-26654-1.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. Gromov–Wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2664–2672. PMLR, 2016.

Moritz Piening and Robert Beinert. A novel sliced fused Gromov-Wasserstein distance. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(29):24828–24836, 2026. doi: 10.1609/aaai.v40i29.39669.

Friedrich Pukelsheim. *Optimal Design of Experiments*, volume 50 of *Classics in Applied Mathematics*. Society for Industrial and Applied Mathematics, 2006. ISBN 9780898716047. doi: 10.1137/1.9780898719109. URL `https://epubs.siam.org/doi/abs/10.1137/1.9780898719109`.

Dan Raviv, Anastasia Dubrovina, and Ron Kimmel. Hierarchical matching of non-rigid shapes. In *Scale Space and Variational Methods in Computer Vision (SSVM 2011)*, volume 6667 of *Lecture Notes in Computer Science*, pages 604–615. Springer, 2012. doi: 10.1007/978-3-642-24785-9_51.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685. IEEE, 2022. doi: 10.1109/cvpr52688.2022.01042.

Ryoma Sato, Marco Cuturi, Makoto Yamada, and Hisashi Kashima. Fast and robust comparison of probability measures in heterogeneous spaces. *arXiv preprint arXiv:2002.01615*, 2020.

Dominik Schnaus, Nikita Araslanov, and Daniel Cremers. It's a (blind) match! Towards vision-language correspondence without parallel data. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24983–24992, 2025. doi: 10.1109/cvpr52734.2025.02326.

Dominik Schnaus, Thomas Dagès, Daniel Cremers, Xi Wang, and Phillip Isola. Shared geometry as a Rosetta Stone: Cross-modal alignment without paired data. arXiv preprint arXiv:2610.09411, 2026. URL `https://arxiv.org/abs/2610.09411`.

Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021. doi: 10.1109/JPROC.2021.3058954. URL `https://doi.org/10.1109/JPROC.2021.3058954`.

Peter H. Schönemann. A generalized solution of the orthogonal Procrustes problem. *Psychometrika*, 31:1–10, 1966. doi: 10.1007/bf02289451.

Athina Sotiropoulou and David Alvarez-Melis. Strongly isomorphic neural optimal transport across incomparable spaces. In *Proceedings of the Geometry-grounded Representation Learning and Generative Modeling Workshop (GRaM)*, volume 251 of *Proceedings of Machine Learning Research*, pages 381–393. PMLR, 2024. doi: 10.48550/arxiv.2407.14957.

Karl-Theodor Sturm. On the geometry of metric measure spaces. *Acta Mathematica*, 196:65–131, 2006. doi: 10.1007/s11511-006-0002-8.

Stas Syrota, Pablo Moreno-Muñoz, and Søren Hauberg. Decoder ensembling for learned latent geometries. In *Proceedings of the Geometry-grounded Representation Learning and Generative Modeling Workshop (GRaM)*, volume 251 of *Proceedings of Machine Learning Research*, pages 277–285. PMLR, 2024. doi: 10.48550/arxiv.2408.07507.

Stas Syrota, Yevgen Zainchkovskyy, Johnny Xi, Benjamin Bloem-Reddy, and Søren Hauberg. Identifying metric structures of deep latent variable models. In *International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 58065–58087. PMLR, 2025. doi: 10.48550/arxiv.2502.13757.

The Tabula Sapiens Consortium. The Tabula Sapiens: A multiple-organ, single-cell transcriptomic atlas of humans. *Science*, 376(6594):eabl4896, 2022. doi: 10.1126/science.abl4896.

Jakub M. Tomczak and Max Welling. VAE with a VampPrior. In Amos Storkey and Fernando Perez-Cruz, editors, *International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1214–1223. PMLR, 2018. doi: 10.48550/arxiv.1705.07120. URL `https://proceedings.mlr.press/v84/tomczak18a.html`.

Théo Uscidda, Luca Eyring, Karsten Roth, Fabian Theis, Zeynep Akata, and Marco Cuturi. Disentangled representation learning with the Gromov-Monge gap. In *International Conference on Learning Representations*, 2025. doi: 10.48550/arxiv.2407.07829.

Titouan Vayer, Laetitia Chapel, Rémi Flamary, Romain Tavenard, and Nicolas Courty. Optimal transport for structured data with application on graphs. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *PMLR*, pages 6275–6284, 2019.

Cédric Villani. *Optimal Transport: Old and New*, volume 338 of *Grundlehren der mathematischen Wissenschaften*. Springer, 2009. ISBN 9783540710509. doi: 10.1007/978-3-540-71050-9.

Tao Wang and Ziv Goldfeld. Neural entropic optimal transport and Gromov-Wasserstein alignment. *IEEE Transactions on Information Theory*, 72:2424–2443, 2026. doi: 10.1109/tit.2026.3661439.

Tongzhou Wang and Phillip Isola. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 9929–9939. PMLR, 2020.

Quanhan Xi and Benjamin Bloem-Reddy. Indeterminacy in generative models: Characterization and strong identifiability. In *International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pages 6912–6939. PMLR, 2023. doi: 10.48550/arxiv.2206.00801.

Han Xiao, Kashif Rasul, and Roland Vollgraf. Fashion-MNIST: a novel image dataset for benchmarking machine learning algorithms. *arXiv preprint arXiv:1708.07747*, 2017.

Hanlin Yu, Berfin Inal, Georgios Arvanitidis, Søren Hauberg, Francesco Locatello, and Marco Fumero. Connecting neural models latent geometries with relative geodesic representations. In *Advances in Neural Information Processing Systems*, volume 38, pages 125316–125360, 2025. doi: 10.52202/085713-3769.

Hongyi Zhang, Yann N. Dauphin, and Tengyu Ma. Fixup initialization: Residual learning without normalization. In *International Conference on Learning Representations*, 2019. doi: 10.48550/arxiv.1901.09321.

Jun-Yan Zhu, Taesung Park, Phillip Isola, and Alexei A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. In *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 2242–2251, 2017. ISBN 9781538610329. doi: 10.1109/iccv.2017.244.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The paper's opening presents the key theoretical finding. The introduction also provides a bit more detail than usual, but references to relevant sections in the paper are included.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Section 8 states them: our recovery guarantees hold only for global optima, which NGMA's solver may not reach, and NGMA's maps act on whole latent spaces of equal dimension, although two models can choose different latent dimensions. The assumptions of the theory, H1–H2 and C1–C2, are stated where they are used.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The paper is predominantly theoretical, and the assumptions are key to understanding the work. We have placed notable effort into keeping assumptions clear and understandable. Complete proofs are in Appendices B and C.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The appendix details experiments, including algorithmic descriptions. We also provide full source code to reproduce all experiments in `https://github.com/mustass/identifiable-alignment`.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We release the code at the URL in Section 5. All experiments rely on either synthetic data (which can be reproduced by the code) or standard public benchmark data.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 6 and Appendix E give the data splits, models, sample sizes, hyperparameters, and each baseline and how it is run.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Tables and figures report the mean $\pm$ s.d. over model pairings and draws, or the median and range over seeds, as each caption states. Table 7 reports means only, and the comparison of segments with geodesics (Appendix E.4) uses one model pair.

   Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Appendix E.7 gives the device, memory and time of each stage of the image experiments, and the total compute of the project. The scVI experiments need far less compute.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research conforms to the NeurIPS Code of Ethics. It uses only public datasets under their terms, including CelebA face images and Tabula Sapiens donor data, and collects no new data.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: Advances in this area are known to have effects for database security [Jha et al., 2025] and our method relies on potentially computationally intensive machinery contribution to high energy consumption.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: Not applicable.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: Every dataset is cited where it is used, with its terms: MNIST [LeCun et al., 1998] (CC BY-SA 3.0), FashionMNIST [Xiao et al., 2017] (MIT licence), CelebA [Liu et al., 2015] (non-commercial research use only, without redistribution) and the Tabula Sapiens lung data [The Tabula Sapiens Consortium, 2022] (Tabula Sapiens data release policy). We use scvi-tools [Lopez et al., 2018] (BSD 3-Clause) and POT [Flamary et al., 2021] (MIT licence), and run Rosetta and mini-vec2vec with the code released by their authors. Every baseline is cited where it is used (Appendix E.2).

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We release the code for NGMA and all experiments at the URL in Section 5.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: The paper does not consider human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: The paper does not consider human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: Not applicable.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.

A Background

This appendix collects the mathematical background required to follow the proofs in Appendix B.

A.1 Riemannian manifolds

Throughout, manifolds are compact, connected and of finite dimension d, and may have a boundary, as in H1. Since $\mathcal{M}$ is smooth, the boundary is that of a smooth manifold with boundary. The proofs use this regularity in two places, for the inward-pointing vector field in the proof of Lemma A.11 and for $\mathrm{dVol}_g(\partial\mathcal{M}) = 0$. We write $\operatorname{int}\mathcal{M}$ for the interior, which is a connected Riemannian manifold without boundary that need not be complete. Statements below about manifolds without boundary are applied to $\operatorname{int}\mathcal{M}$. Proofs and a broader exposition are in do Carmo [1992], Petersen [2016], Lee [2018].

Definition A.1 (do Carmo, 1992, Lee, 2018)**.** A *Riemannian manifold* $(\mathcal{M}, g)$ is a smooth manifold $\mathcal{M}$ equipped with a smoothly varying family of inner products g_x on each tangent space $T_x\mathcal{M}$. The metric g induces:

- the *geodesic distance* $d_g(x, y) = \inf_\gamma \int_0^1 \sqrt{g_{\gamma(t)}(\dot\gamma(t), \dot\gamma(t))}\, dt$, where the infimum is over piecewise smooth curves joining x to y;
- the *Riemannian volume measure* dVol_g, the unique measure that in any chart equals $\sqrt{\det(g_{ij})}\, dx^1 \cdots dx^d$;

Definition A.2 (Lee, 2018)**.** For each $x \in \mathcal{M}$ the *exponential map* $\exp_x : T_x\mathcal{M} \to \mathcal{M}$ sends a tangent vector v to the endpoint $\gamma(1)$ of the geodesic γ with $\gamma(0) = x$ and $\dot\gamma(0) = v$.

Remark A.3 (do Carmo, 1992)*.* On a compact manifold without boundary $\exp_x$ is defined on all of $T_x\mathcal{M}$. With a boundary it is defined near the origin of $T_x\mathcal{M}$ for interior x, which is all that Corollary A.7 uses.

Definition A.4 (Kobayashi and Nomizu, 1996, Petersen, 2016)**.** A *Riemannian isometry* $T : (\mathcal{M}, g) \to (\mathcal{N}, h)$ is a diffeomorphism satisfying the pullback condition $T^*h = g$, i.e. $g_x(u, v) = h_{T(x)}(dT_x\, u,\, dT_x\, v)$ for all $x \in \mathcal{M}$ and $u, v \in T_x\mathcal{M}$. Isometries preserve all metric quantities:

- geodesic distances: $d_h(T(x), T(y)) = d_g(x, y)$;
- metric balls: $T\big(\bar B_{\mathcal{M}}(x, r)\big) = \bar B_{\mathcal{N}}(T(x), r)$ for every $x \in \mathcal{M}$ and $r \geq 0$, where $\bar B_{\mathcal{M}}(x, r)$ is the closed ball of radius r around x;
- the differential at each point is an orthogonal linear map: $dT_x \in O(T_x\mathcal{M}, g_x)$.

Theorem A.5 (Myers–Steenrod; Myers and Steenrod, 1939, Palais, 1957)**.** *A distance-preserving bijection of a connected Riemannian manifold without boundary, complete or not, is a smooth Riemannian isometry. The isometry group* $\mathrm{Iso}(\mathcal{M}, g)$ *of such a manifold is a Lie group with respect to the compact-open topology, acting smoothly and effectively on* $\mathcal{M}$*. If* $\mathcal{M}$ *is compact, then* $\mathrm{Iso}(\mathcal{M}, g)$ *is compact, and in particular the component group* $\pi_0(\mathrm{Iso}(\mathcal{M}, g))$ *is finite.*

Proposition A.6 (1-jet rigidity; Kobayashi and Nomizu, 1996, Myers and Steenrod, 1939)**.** *A self-isometry of a connected Riemannian manifold, complete or not, is uniquely determined by its value and differential at a single point. Concretely, if* $\varphi : \mathcal{M} \to \mathcal{M}$ *is an isometry with* $\varphi(x_0) = x_0$ *and* $d_{x_0}\varphi = \mathrm{id}$*, then* $\varphi = \mathrm{id}_{\mathcal{M}}$*.*

Corollary A.7 (Pinning an isometry by $d+1$ points)**.** *Let* φ *be an isometry of* $\operatorname{int}\mathcal{M}$*, let* $x_1 \in \operatorname{int}\mathcal{M}$*, and let* $B = \exp_{x_1}(\{v : |v|_g < \varepsilon\}) \subset \operatorname{int}\mathcal{M}$ *be a normal ball, so that* $\exp_{x_1}$ *is a diffeomorphism onto* B *and each* $y \in B$ *is joined to* x_1 *by a unique geodesic of length less than* ε*. If* φ *fixes* x_1 *and points* $x_2, \dots, x_{d+1} \in B$ *whose log-map images* $v_j = \exp_{x_1}^{-1}(x_j)$ *are linearly independent, then* $\varphi = \mathrm{id}$*, and so is its extension to* $\mathcal{M}$*.*

Proof. φ fixes x_1 and maps the geodesic $t \mapsto \exp_{x_1}(tv_j)$, $t \in [0, 1]$, to a geodesic from x_1 to x_j of the same length, which is less than ε. By uniqueness it is the same geodesic, so $d_{x_1}\varphi\, v_j = v_j$ for $j = 2, \dots, d{+}1$. An orthogonal map fixing a basis of $T_{x_1}\mathcal{M}$ is the identity, so $d_{x_1}\varphi = \mathrm{id}$ and 1-jet rigidity on the connected manifold $\operatorname{int}\mathcal{M}$ gives $\varphi = \mathrm{id}$. The extension to $\mathcal{M}$ is the identity by continuity, since $\operatorname{int}\mathcal{M}$ is dense. □

Remark A.8. If $x_2, \dots, x_{d+1}$ are drawn independently from an absolutely continuous measure restricted to B, the log-map images are linearly independent almost surely. Conditionally on $x_2, \dots, x_j$, the bad set for x_{j+1} is the image under $\exp_{x_1}$ of a proper linear subspace, which is dVol_g-null. The condition is local, since it needs $d+1$ points in general position close together, not $d+1$ arbitrary points.

A.2 Metric measure spaces

Definition A.9 (Gromov, 1999, Mémoli, 2011)**.** A *metric measure space* (mm-space) $(\mathcal{M}, d_{\mathcal{M}}, \mu)$ is a compact metric space with a Borel probability measure. Two mm-spaces are *isomorphic* if a bijection $T\colon \mathcal{M} \to \mathcal{N}$ preserves distances and satisfies $T_{\#}\mu = \nu$; we write $\mathrm{Iso}\big((\mathcal{M}, d_{\mathcal{M}}, \mu), (\mathcal{N}, d_{\mathcal{N}}, \nu)\big)$ for the set of such maps. The *automorphism group* is

$$\mathrm{Aut}(\mathcal{M}, d_{\mathcal{M}}, \mu) \;=\; \mathrm{Iso}\big((\mathcal{M}, d_{\mathcal{M}}, \mu),\, (\mathcal{M}, d_{\mathcal{M}}, \mu)\big),$$

the distance-preserving bijections of $\mathcal{M}$ that preserve μ. It forms a group under composition.

Remark A.10 (Notation). In the appendices $(\mathcal{M}, g, \mu)$ denotes the mm-space $(\mathcal{M}, d_g, \mu)$, where d_g is the geodesic distance of the Riemannian metric g defined above. The data mm-space $(\mathcal{M}, d_{\mathcal{M}}, P_{\mathcal{M}})$ and the latent mm-spaces $(\mathcal{Z}_a, d_a, P_a)$ of Section 2 are of this form. Statements proved for a general $(\mathcal{M}, g, \mu)$ apply to each of them, with C1 and C2 read as conditions on that space; the proof of Proposition B.7 shows that their automorphism groups are conjugate.

Lemma A.11 (Distance isometries are Riemannian isometries)**.** *Let $(\mathcal{M}, g)$ be a compact connected Riemannian manifold, possibly with a boundary, with geodesic distance d_g. Restriction to $\mathrm{int}\,\mathcal{M}$ is a bijection from the distance-preserving bijections of $(\mathcal{M}, d_g)$ onto the Riemannian isometries of $(\mathrm{int}\,\mathcal{M}, g)$, and this group is a compact Lie group acting smoothly and effectively on $\mathrm{int}\,\mathcal{M}$. If $\mu = \rho\,\mathrm{dVol}_g$ with ρ continuous, a distance-preserving bijection φ satisfies $\varphi_{\#}\mu = \mu$ if and only if $\rho \circ \varphi = \rho$.*

Proof. Let $d_\circ$ be the length metric of $\mathrm{int}\,\mathcal{M}$. Clearly $d_g \le d_\circ$ on $\mathrm{int}\,\mathcal{M}$. For the reverse inequality, let X be a smooth vector field on $\mathcal{M}$ that points strictly inward along $\partial\mathcal{M}$, with flow θ_s. For small $s > 0$, θ_s maps $\mathcal{M}$ into $\mathrm{int}\,\mathcal{M}$, and $\theta_s \to \mathrm{id}$ in C^1 as $s \to 0$. For a curve γ in $\mathcal{M}$ between interior points x and y, the curve that follows the flow line from x to $\theta_s(x)$, then $\theta_s \circ \gamma$, then the flow line from $\theta_s(y)$ back to y, stays in $\mathrm{int}\,\mathcal{M}$ and has length at most $(1 + o(1))\,L(\gamma) + O(s)$. Hence $d_\circ = d_g$ on $\mathrm{int}\,\mathcal{M}$.

A distance-preserving bijection φ of $\mathcal{M}$ is a homeomorphism, so by invariance of domain it maps $\mathrm{int}\,\mathcal{M}$ onto itself. Its restriction preserves $d_\circ$, and is therefore a smooth Riemannian isometry by the Myers–Steenrod theorem (Theorem A.5), which does not require completeness. Conversely, a Riemannian isometry of $\mathrm{int}\,\mathcal{M}$ preserves lengths of curves, hence $d_\circ = d_g$, and extends uniquely to a distance-preserving bijection of $\mathcal{M}$, since $\mathrm{int}\,\mathcal{M}$ is dense in the compact space $\mathcal{M}$. The restriction map is a bijection and a group homomorphism. The group of distance-preserving bijections of a compact metric space is compact in the uniform topology by the Arzelà–Ascoli theorem, and the restriction map is a continuous bijection from it onto $\mathrm{Iso}(\mathrm{int}\,\mathcal{M}, g)$ with the compact-open topology, hence a homeomorphism. By Theorem A.5 the latter is a Lie group acting smoothly and effectively.

Finally, $\mu(\partial\mathcal{M}) = 0$, and an isometry preserves dVol_g, so $\varphi_{\#}\mu = (\rho \circ \varphi^{-1})\,\mathrm{dVol}_g$. This equals μ if and only if $\rho \circ \varphi^{-1} = \rho$ almost everywhere, hence everywhere by continuity, which is equivalent to $\rho \circ \varphi = \rho$. □

Remark A.12. The following two identities, used repeatedly in the proofs, follow from the chain rule for pushforwards:

$$(T \circ \phi)_{\#}\mu = T_{\#}(\phi_{\#}\mu),$$
$$(T^{-1})_{\#}\nu = (T_{\#})^{-1}\nu.$$

In particular, if $\phi \in \mathrm{Aut}(\mathcal{M}, g, \mu)$ and $T \in \mathrm{Iso}\big((\mathcal{M}, g, \mu), (\mathcal{N}, h, \nu)\big)$, then $T \circ \phi \in \mathrm{Iso}\big((\mathcal{M}, g, \mu), (\mathcal{N}, h, \nu)\big)$ because $(T \circ \phi)_{\#}\mu = T_{\#}(\phi_{\#}\mu) = T_{\#}\mu = \nu$.

For an mm-space $(\mathcal{M}, d_{\mathcal{M}}, \mu)$, a point $x \in \mathcal{M}$ and $r \ge 0$, the *local distance distribution* (LDD) of x is $\Phi_{\mathcal{M}}(x, r) := \mu\big(\bar{B}_{\mathcal{M}}(x, r)\big)$, as in Section 4.1.

Proposition A.13 (Mémoli, 2011, Mémoli and Needham, 2022)**.** *LDDs are preserved by every isomorphism $T : (\mathcal{M}, g, \mu) \to (\mathcal{N}, h, \nu)$:*

$$\Phi_{\mathcal{N}}(T(x), r) = \Phi_{\mathcal{M}}(x, r) \qquad \forall\, x \in \mathcal{M},\ r \geq 0.$$

Remark A.14. This follows from two properties of T:

1. *Isometry maps balls to balls of the same radius:* $T\big(\bar{B}_{\mathcal{M}}(x, r)\big) = \bar{B}_{\mathcal{N}}(T(x), r)$, since T preserves geodesic distances.
2. *Measure preservation:* $\nu(T(S)) = \mu(S)$ for every measurable $S \subseteq \mathcal{M}$, since $T_{\#}\mu = \nu$.

Combining the two, $\Phi_{\mathcal{N}}(T(x), r) = \nu\big(\bar{B}_{\mathcal{N}}(T(x), r)\big) = \nu\big(T(\bar{B}_{\mathcal{M}}(x, r))\big) = \mu\big(\bar{B}_{\mathcal{M}}(x, r)\big) = \Phi_{\mathcal{M}}(x, r)$.

A.3 Analysis and optimal transport

Proposition A.15 (Coarea formula; Chavel, 1984, Evans and Gariepy, 2015)**.** *Let $f : \mathcal{M} \to \mathbb{R}$ be a Lipschitz function and $u : \mathcal{M} \to \mathbb{R}$ integrable. Then*

$$\int_{\mathcal{M}} u(x)\, |\nabla f(x)|\ \mathrm{dVol}_g \ = \ \int_{-\infty}^{\infty} \int_{f^{-1}(t)} u\, d\mathcal{H}^{d-1}\, dt,$$

where $\mathcal{H}^{d-1}$ is the $(d-1)$-dimensional Hausdorff measure.

Remark A.16. A consequence used in the proofs of Appendix B is that distance spheres have measure zero. For any $p \in \mathcal{M}$ and $c > 0$, the distance function $z \mapsto d_g(z, p)$ has $|\nabla d_g(\cdot, p)| = 1$ almost everywhere [do Carmo, 1992], so the coarea formula gives $\mathrm{Vol}_g\big(\{z : d_g(z, p) = c\}\big) = 0$. The same holds with a boundary. The function $d_g(\cdot, p)$ is 1-Lipschitz, and for interior $z \neq p$ a shortest curve from p to z exists by compactness and ends in a geodesic segment of $\operatorname{int}\mathcal{M}$, along which $d_g(\cdot, p)$ grows at unit rate, so $|\nabla d_g(\cdot, p)| = 1$ wherever it is differentiable, which is almost everywhere; the boundary is Vol_g-null.

Appendix C uses one more classical fact.

Lemma A.17 (Zero sets of real-analytic functions; Mityagin, 2020, Krantz and Parks, 2002)**.** *Let $\Omega \subseteq \mathbb{R}^n$ be open and connected and $f : \Omega \to \mathbb{R}$ real-analytic with $f \not\equiv 0$. Then the zero set $\{z \in \Omega : f(z) = 0\}$ has Lebesgue measure zero. The same holds, chartwise, on any connected real-analytic manifold.*

We recall some standard facts from optimal transport [Villani, 2009].

Definition A.18. Given probability measures μ on $\mathcal{M}$ and ν on $\mathcal{N}$, the set of *couplings* is

$$\Pi(\mu, \nu) = \big\{\eta \in \mathcal{P}(\mathcal{M} \times \mathcal{N}) \ \big|\ (\pi_{\mathcal{M}})_{\#}\eta = \mu,\ (\pi_{\mathcal{N}})_{\#}\eta = \nu\big\},$$

where $\pi_{\mathcal{M}}$ and $\pi_{\mathcal{N}}$ are the coordinate projections.

Proposition A.19 (Couplings; Villani, 2009)**.**

- *$\Pi(\mu, \nu)$ is tight (its members have the fixed marginals μ and ν on compact spaces), so by Prokhorov's theorem every sequence of couplings has a weakly convergent subsequence.*
- *Transport cost functionals of the form $\eta \mapsto \int c\, d\eta$ with continuous c are weakly lower semicontinuous, so minimisers exist.*
- *A coupling η is the* graph measure *of a measurable map $T : \mathcal{M} \to \mathcal{N}$ if $\eta = (\mathrm{id}_{\mathcal{M}} \times T)_{\#}\mu$; equivalently, η is concentrated on the graph $\{(x, T(x)) : x \in \mathcal{M}\}$. In this case, η-almost every pair (x, y) satisfies $y = T(x)$.*

We also recall a fact from measure theory [Evans and Gariepy, 2015].

Proposition A.20. *The support $\operatorname{supp}(\eta) \subseteq \mathcal{M} \times \mathcal{N}$ is the smallest closed set of full η-measure. If μ and ν have full support (i.e. $\operatorname{supp}(\mu) = \mathcal{M}$ and $\operatorname{supp}(\nu) = \mathcal{N}$, which holds when $\mu = \rho\, \mathrm{dVol}_g$ with density $\rho > 0$ everywhere and likewise for ν, as for P_a and P_b), then for any $\eta \in \Pi(\mu, \nu)$:*

$$\pi_{\mathcal{M}}\big(\operatorname{supp}(\eta)\big) = \mathcal{M}, \qquad \pi_{\mathcal{N}}\big(\operatorname{supp}(\eta)\big) = \mathcal{N}.$$

Remark A.21. This surjectivity of the projections is used in the proof of Proposition 4.1 to upgrade a zero-distortion condition on a subset $S \subseteq \mathcal{M} \times \mathcal{N}$ to a correspondence covering all of $\mathcal{M}$ and $\mathcal{N}$.

The margin test of Section 4.1 rests on the following concentration inequality for empirical distribution functions.

Theorem A.22 (Dvoretzky–Kiefer–Wolfowitz inequality, with Massart's constant; Massart, 1990)**.** *Let F be a distribution function on $\mathbb{R}$ and let $\widehat{F}_N$ be the empirical distribution function of N i.i.d. draws from F. For every $\varepsilon > 0$,*

$$P\Big(\sup_{t\in\mathbb{R}} \big|\widehat{F}_N(t) - F(t)\big| > \varepsilon\Big) \;\le\; 2e^{-2N\varepsilon^2}.$$

The genericity result of Appendix B uses the following topology on smooth functions.

Definition A.23 (Whitney C^∞ topology)**.** Fix a finite atlas (U_i, φ_i) of $\mathcal{M}$ and compact sets $K_i \subset U_i$ that cover $\mathcal{M}$. The *Whitney C^∞ topology* on $C^\infty(\mathcal{M})$ is defined by the family of seminorms $\|\rho\|_{C^k} = \max_i \sup_{|\alpha|\le k} \sup_{\varphi_i(K_i)} |\partial^\alpha(\rho \circ \varphi_i^{-1})|$ for $k = 0, 1, 2, \dots$ (any two choices of atlas and compact sets give equivalent seminorms on a compact manifold). A sequence $\rho_n \to \rho$ iff all partial derivatives converge uniformly. This makes $C^\infty(\mathcal{M})$ a Fréchet space (a complete, metrisable, locally convex topological vector space). The subset $C^\infty_{>0}(\mathcal{M}) = \{\rho \in C^\infty(\mathcal{M}) \mid \rho > 0\}$ is open in this topology.

A.4 Lie groups and Lie algebras

The following are general results and definitions from Hall [2015], Lee [2012].

Definition A.24 (Lie group)**.** A *Lie group* is a smooth manifold G that is also a group, such that the multiplication $(g, h) \mapsto gh$ and inversion $g \mapsto g^{-1}$ are both smooth maps.

Example A.25. The group $O(n) = \{Q \in \mathbb{R}^{n\times n} \mid Q^\top Q = I\}$ of orthogonal matrices is a compact Lie group of dimension $\frac{n(n-1)}{2}$. The subgroup $SO(n) = \{Q \in O(n) \mid \det Q = 1\}$ of rotations is also a compact Lie group.

Definition A.26 (Lie algebra of a Lie group)**.** The *Lie algebra* $\mathfrak{g}$ of a Lie group G is the vector space T_eG (the tangent space at the identity) equipped with a bracket $[\cdot,\cdot] : \mathfrak{g} \times \mathfrak{g} \to \mathfrak{g}$ that is bilinear, antisymmetric ($[X, Y] = -[Y, X]$), and satisfies the Jacobi identity. Concretely, elements of $\mathfrak{g}$ are identified with left-invariant vector fields on G, and the bracket is the usual Lie bracket of vector fields.

The dimension of $\mathfrak{g}$ as a vector space equals the dimension of G as a manifold.

Example A.27. The Lie algebra $\mathfrak{o}(3)$ of $O(3)$ is the 3-dimensional space of skew-symmetric matrices $\{X \in \mathbb{R}^{3\times 3} \mid X^\top = -X\}$, with bracket $[X, Y] = XY - YX$. A standard basis is:

$$X_1 = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -1 \\ 0 & 1 & 0 \end{pmatrix}, \quad X_2 = \begin{pmatrix} 0 & 0 & 1 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \end{pmatrix}, \quad X_3 = \begin{pmatrix} 0 & -1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix},$$

generating rotations about the x-, y-, and z-axes respectively.

Definition A.28 (One-parameter subgroup)**.** A *one-parameter subgroup* of a Lie group G is a smooth homomorphism $\gamma : (\mathbb{R}, +) \to G$, i.e., a smooth curve satisfying $\gamma(s + t) = \gamma(s)\gamma(t)$ for all $s, t \in \mathbb{R}$ and $\gamma(0) = e$.

Definition A.29 (Exponential map)**.** The *exponential map* $\exp : \mathfrak{g} \to G$ sends each $X \in \mathfrak{g}$ to $\gamma_X(1)$, where γ_X is the unique one-parameter subgroup with $\dot\gamma_X(0) = X$. The full one-parameter subgroup is then $t \mapsto \exp(tX)$.

Example A.30. In $O(3)$, the one-parameter subgroup generated by X_3 is the family of rotation matrices:

$$\exp(tX_3) = \begin{pmatrix} \cos t & -\sin t & 0 \\ \sin t & \cos t & 0 \\ 0 & 0 & 1 \end{pmatrix},$$

which rotates about the z-axis by angle t.

Definition A.31 (Smooth group action)**.** A *smooth (left) action* of a Lie group G on a smooth manifold $\mathcal{M}$ is a smooth map $\alpha : G \times \mathcal{M} \to \mathcal{M}$, written $\alpha(g, p) = g \cdot p$, satisfying $e \cdot p = p$ and $(gh) \cdot p = g \cdot (h \cdot p)$ for all $g, h \in G$ and $p \in \mathcal{M}$.

Definition A.32 (Infinitesimal generator)**.** Given a smooth action of G on $\mathcal{M}$ and an element $X \in \mathfrak{g}$, the *infinitesimal generator* $X^\sharp$ is the vector field on $\mathcal{M}$ defined by

$$X^\sharp_p = \left.\frac{d}{dt}\right|_{t=0} \exp(tX) \cdot p.$$

It is the velocity field on $\mathcal{M}$ induced by the flow $t \mapsto \exp(tX) \cdot p$.

Definition A.33 (Effective action)**.** An action is *effective* (or *faithful*) if the only group element that fixes every point is the identity: $g \cdot p = p$ for all $p \in \mathcal{M}$ implies $g = e$.

Example A.34. The standard action of $O(3)$ on S^2 is effective. The infinitesimal generator of X_3 (rotation about the z-axis) is the vector field $X_3^\sharp = \partial_\phi$ in spherical coordinates (θ, ϕ). It points eastward at every point, with magnitude $\sin\theta$.

Definition A.35 (Lie derivative)**.** Let X be a vector field on $\mathcal{M}$ with flow φ_t. The *Lie derivative* of a smooth function $\rho : \mathcal{M} \to \mathbb{R}$ along X is:

$$L_X\rho(p) = \left.\frac{d}{dt}\right|_{t=0} \rho(\varphi_t(p)) = d\rho_p(X_p).$$

In local coordinates, $L_X\rho = X^i\partial_i\rho$, the directional derivative of ρ along X.

The following fact converts a global symmetry condition into a differential equation.

Lemma A.36 (Invariance via Lie derivative)**.** *Let G act smoothly on $\mathcal{M}$, let $X \in \mathfrak{g}$, and let $\rho \in C^\infty(\mathcal{M})$. Then ρ is invariant under the one-parameter subgroup $\{\exp(tX)\}_{t\in\mathbb{R}}$, meaning $\rho(\exp(tX) \cdot p) = \rho(p)$ for all t and all p, if and only if $L_{X^\sharp}\rho = 0$ everywhere on $\mathcal{M}$.*

Proof. Define $h(t) = \rho(\exp(tX)\cdot p)$. Then $h'(t) = (L_{X^\sharp}\rho)(\exp(tX)\cdot p)$. If $L_{X^\sharp}\rho = 0$, then $h' \equiv 0$, so $h(t) = h(0) = \rho(p)$. Conversely, if $h(t) = h(0)$ for all t and p, then $h'(0) = L_{X^\sharp}\rho(p) = 0$. □

Example A.37. On S^2, the condition $L_{X_3^\sharp}\rho = \partial_\phi\rho = 0$ says that ρ does not depend on longitude, that is, ρ is rotationally symmetric about the z-axis. Imposing all three conditions $L_{X_1^\sharp}\rho = L_{X_2^\sharp}\rho = L_{X_3^\sharp}\rho = 0$ forces ρ to be constant on S^2.

Definition A.38 (Identity component)**.** The *identity component* G_0 of a Lie group G is the connected component containing the identity element e. It is a closed normal subgroup of G with $\dim G_0 = \dim G$ and the same Lie algebra $\mathfrak{g}$.

Definition A.39 (Component group)**.** The quotient group $\pi_0(G) := G/G_0$ is called the *component group*. Its elements are the connected components of G, and it is a discrete group.

Example A.40. For $O(3)$, $G_0 = SO(3)$ (proper rotations), and $\pi_0(O(3)) \cong \mathbb{Z}/2\mathbb{Z}$ with the non-trivial element represented by any orientation-reversing isometry (e.g., the antipodal map $p \mapsto -p$).

B Proofs

Throughout the appendix, $d := \dim \mathcal{M}$.

B.1 Proofs for Section 2

Remark B.1 (The projection onto $\mathcal{M}$). Here π is the nearest-point map in the Euclidean distance of $\mathcal{X}$. The distance to the closed set $\mathcal{M}$ is 1-Lipschitz [Federer, 1959, 4.8(1)], hence differentiable almost everywhere by Rademacher's theorem, and every point where it is differentiable has a unique nearest point on $\mathcal{M}$ [Federer, 1959, 4.8(3)], so π is defined Lebesgue-almost everywhere and is continuous on its domain [Federer, 1959, 4.8(4)]. Hence $\pi_\# P_\mathcal{X}$ is well defined whenever $P_\mathcal{X}$ has a density. The noise condition of H1, normal to $\mathcal{M}$ and shorter than the reach [Federer, 1959], is what makes $\pi(x)$ the point that x was generated from, so that $\pi_\# P_\mathcal{X}$ is the distribution of the generating points and not a blurred version of it; within the reach π is moreover Lipschitz [Federer, 1959, 4.8(8)]. The proofs

use only that $P_{\mathcal{M}}$ has a smooth positive density on $\mathcal{M}$. That assumption also rules out observations beyond the boundary of $\mathcal{M}$, which would project onto $\partial\mathcal{M}$, a set of volume zero; normal noise never produces them.

Lemma B.2 (Sufficient conditions for H2)**.** *Let $\mathcal{M} \subseteq \mathcal{X}$ be a compact d-dimensional embedded submanifold, possibly with smooth boundary.* (i) *Let $\mathcal{W}$ be a manifold of dimension $q \geq d$, such as $\mathbb{R}^q$, and $\mathrm{dec}\colon \mathcal{W} \to \mathcal{X}$ smooth and injective with a Jacobian of rank q, with $\mathcal{M} \subseteq \mathrm{dec}(\mathcal{W})$ and $\mathcal{Z} := \mathrm{dec}^{-1}(\mathcal{M})$ compact. Then $\mathcal{Z}$ is a compact embedded submanifold of $\mathcal{W}$ of dimension d and $\mathrm{dec}\,|_{\mathcal{Z}}\colon \mathcal{Z} \to \mathcal{M}$ is a diffeomorphism, so $\mathrm{dec}\,|_{\mathcal{Z}}$ satisfies H2. The preimage is closed, so it is compact whenever $\mathcal{W}$ is, and then dec is a smooth embedding [Syrota et al., 2025, Lemma 4.1]; it is also compact whenever dec is a smooth embedding.* (ii) *Let $\mathrm{enc}\colon \mathcal{X} \to \mathbb{R}^q$ be smooth, injective on $\mathcal{M}$, with a Jacobian of rank d on the tangent spaces of $\mathcal{M}$. Then $\mathrm{enc}(\mathcal{M})$ is a compact embedded submanifold of $\mathbb{R}^q$ and $(\mathrm{enc}\,|_{\mathcal{M}})^{-1}\colon \mathrm{enc}(\mathcal{M}) \to \mathcal{M}$ satisfies H2.*

Proof. (i) Suppose first that dec is an embedding, so $\mathcal{N} := \mathrm{dec}(\mathcal{W})$ is an embedded q-dimensional submanifold of $\mathcal{X}$ and $\mathrm{dec}\colon \mathcal{W} \to \mathcal{N}$ is a diffeomorphism. The inclusion $\mathcal{M} \hookrightarrow \mathcal{X}$ has image in $\mathcal{N}$, so it is smooth as a map into $\mathcal{N}$ [Lee, 2018], and it is an injective immersion with compact domain, hence an embedding. Composing with $\mathrm{dec}^{-1}\colon \mathcal{N} \to \mathcal{W}$ shows that $\mathrm{dec}^{-1}\,|_{\mathcal{M}}$ embeds $\mathcal{M}$ into $\mathcal{W}$ with image $\mathcal{Z}$, and $\mathrm{dec}\,|_{\mathcal{Z}}$ is its inverse.

In general, $\mathcal{Z}$ is compact, so it has an open neighbourhood U in $\mathcal{W}$ whose closure $\bar{U}$ is compact. On $\bar{U}$ the map dec is continuous and injective, hence a homeomorphism onto its image, and so is its restriction to U. Thus $\mathrm{dec}\,|_U$ is an injective immersion and a homeomorphism onto its image, that is, a smooth embedding, with $\mathcal{M} \subseteq \mathrm{dec}(U)$ and $U \cap \mathrm{dec}^{-1}(\mathcal{M}) = \mathcal{Z}$. The first paragraph applied to $\mathrm{dec}\,|_U$ concludes.

(ii) The restriction $\mathrm{enc}\,|_{\mathcal{M}}$ is an injective immersion of a compact manifold, hence an embedding [Lee, 2018]. So $\mathrm{enc}(\mathcal{M})$ is a compact embedded submanifold of $\mathbb{R}^q$ of dimension d and $\mathrm{enc}\,|_{\mathcal{M}}\colon \mathcal{M} \to \mathrm{enc}(\mathcal{M})$ is a diffeomorphism, whose inverse is a diffeomorphism onto $\mathcal{M}$. □

Remark B.3 (An injective immersion is not enough). An injective immersion need not have a continuous inverse on its image. The map $f\colon \mathbb{R} \to \mathbb{R}^2$ with $f(t) = (\cos\vartheta(t), \sin\vartheta(t))$, $\vartheta(t) = 2\pi t/(1+t)$ for $t \geq 0$, extended to $t < 0$ by a curve leaving the circle, is an injective immersion whose image contains the unit circle $\mathcal{M}$. Yet $f^{-1}(\mathcal{M}) = [0, \infty)$, and points of $\mathcal{M}$ on either side of $(1, 0)$ have preimages near 0 and near ∞. The pullback length metric on $f^{-1}(\mathcal{M})$ is then not the length metric of $\mathcal{M}$. An injective decoder with full-rank Jacobian whose image contains $\mathcal{M}$ therefore need not satisfy H2 unless $f^{-1}(\mathcal{M})$ is compact, and Proposition 2.3 can fail for it.

Proposition B.4 (Proposition 2.2)**.** *Under H1–H2, metrics d_a on $\mathcal{Z}_a$ and d_b on $\mathcal{Z}_b$ satisfy $d_b(A(z), A(z')) = d_a(z, z')$ for all $z, z' \in \mathcal{Z}_a$ if and only if $d_a(f_a^{-1}(x), f_a^{-1}(y)) = d_b(f_b^{-1}(x), f_b^{-1}(y))$ for all $x, y \in \mathcal{M}$. Euclidean metrics on two subsets of one $\mathbb{R}^q$ satisfy this if and only if A extends to a rigid motion of $\mathbb{R}^q$, and geodesic metrics on two subsets of the unit sphere $S^{q-1} \subset \mathbb{R}^q$ if and only if A extends to an orthogonal map of $\mathbb{R}^q$.*

Proof. Since f_a is a bijection onto $\mathcal{M}$, every pair $z, z' \in \mathcal{Z}_a$ is the pair of representations $z = f_a^{-1}(x)$, $z' = f_a^{-1}(y)$ of exactly one pair $x, y \in \mathcal{M}$, and Definition 2.1 gives $A(z) = f_b^{-1}(x)$ and $A(z') = f_b^{-1}(y)$. Substituting turns one condition into the other.

For the Euclidean case, a rigid motion preserves Euclidean distances, so its restriction does. Conversely, let A preserve them, fix $z_0 \in \mathcal{Z}_a$, and set $u := z - z_0$ and $B(u) := A(z) - A(z_0)$. Then $\|B(u)\| = \|u\|$ and $\|B(u) - B(u')\| = \|u - u'\|$, so the polarisation identity gives $\langle B(u), B(u')\rangle = \langle u, u'\rangle$ for all u, u'. A map that preserves inner products on a set extends to a linear isometry of its span, which extends to an orthogonal map Q of $\mathbb{R}^q$. Hence $A(z) = Q(z - z_0) + A(z_0)$, a rigid motion.

For the spherical case, an orthogonal map preserves inner products, hence the geodesic distance $\arccos\langle z, z'\rangle$ on S^{q-1}. Conversely, if A preserves geodesic distances on the sphere, then $\langle A(z), A(z')\rangle = \cos d(A(z), A(z')) = \cos d(z, z') = \langle z, z'\rangle$, so A preserves inner products and extends to an orthogonal map Q of $\mathbb{R}^q$ as above. The cosine distance is a monotone function of the geodesic distance, so the same holds for it. □

Proposition B.5 (Proposition 2.3)**.** *Under H1–H2 the alignment map exists, is unique, and equals $A = f_b^{-1} \circ f_a$. It satisfies $A^* g_b = g_a$ and $A_\# P_a = P_b$, so it is an isomorphism of $(\mathcal{Z}_a, d_a, P_a)$ and $(\mathcal{Z}_b, d_b, P_b)$.*

Proof. By H2, $f_a|_{\mathcal{Z}_a}$ and $f_b|_{\mathcal{Z}_b}$ are diffeomorphisms onto $\mathcal{M}$. A map A with $f_b \circ A = f_a$ must send z to the unique preimage of $f_a(z)$ under f_b, so $A = f_b^{-1} \circ f_a$ is the only alignment map, and it is a diffeomorphism. The chain rule gives

$$A^* g_b = A^* f_b^* g^{\mathcal{M}} = (f_b \circ A)^* g^{\mathcal{M}} = f_a^* g^{\mathcal{M}} = g_a,$$

as in Syrota et al. [2025, Theorem 4.5]. Hence A maps each piecewise smooth curve in $\mathcal{Z}_a$ to one in $\mathcal{Z}_b$ of equal length, and A^{-1} does the same in reverse, so $d_b(A(z), A(z')) = d_a(z, z')$ for all $z, z' \in \mathcal{Z}_a$. Finally, $A_\# P_a = (f_b^{-1} \circ f_a \circ f_a^{-1})_\# P_{\mathcal{M}} = (f_b^{-1})_\# P_{\mathcal{M}} = P_b$. The same argument with f_a in place of A shows that f_a is an isomorphism from $(\mathcal{Z}_a, d_a, P_a)$ onto $(\mathcal{M}, d_{\mathcal{M}}, P_{\mathcal{M}})$. □

B.2 Proofs for Section 3

Lemma B.6 (Isomorphisms differ by automorphisms)**.** *Let $(\mathcal{M}, g, \mu)$ and $(\mathcal{N}, h, \nu)$ be mm-spaces and $T \in \mathrm{Iso}\big((\mathcal{M}, g, \mu), (\mathcal{N}, h, \nu)\big)$. Then*

$$\mathrm{Iso}\big((\mathcal{M}, g, \mu), (\mathcal{N}, h, \nu)\big) = \{T \circ \phi \,:\, \phi \in \mathrm{Aut}(\mathcal{M}, g, \mu)\}.$$

In particular T is the only isomorphism if and only if $\mathrm{Aut}(\mathcal{M}, g, \mu) = \{\mathrm{id}_{\mathcal{M}}\}$, and then also $\mathrm{Aut}(\mathcal{N}, h, \nu) = \{\mathrm{id}_{\mathcal{N}}\}$, since $\psi \mapsto T^{-1} \circ \psi \circ T$ is a group isomorphism between the two automorphism groups.

Proof. For $\phi \in \mathrm{Aut}(\mathcal{M}, g, \mu)$, the composition $T \circ \phi$ preserves distances and $(T \circ \phi)_\# \mu = T_\#(\phi_\# \mu) = T_\# \mu = \nu$, so it is an isomorphism. Conversely, for an isomorphism T', the map $\phi := T^{-1} \circ T'$ is a distance-preserving bijection of $\mathcal{M}$ with $\phi_\# \mu = (T^{-1})_\# \nu = \mu$, so $\phi \in \mathrm{Aut}(\mathcal{M}, g, \mu)$ and $T' = T \circ \phi$. The map $\phi \mapsto T \circ \phi$ is injective, which gives the second statement. □

Proposition B.7 (Proposition 3.1)**.** *Under H1–H2, the isomorphisms from $(\mathcal{Z}_a, d_a, P_a)$ to $(\mathcal{Z}_b, d_b, P_b)$ are exactly the maps $f_b^{-1} \circ S \circ f_a$ with $S \in \mathrm{Aut}(\mathcal{M}, d_{\mathcal{M}}, P_{\mathcal{M}})$. Hence every such isomorphism equals the alignment A if and only if C1 holds.*

Proof. By Proposition B.5, A is an isomorphism of the latent mm-spaces and f_a is an isomorphism of $(\mathcal{Z}_a, d_a, P_a)$ onto the data mm-space. Conjugation by f_a, $S \mapsto f_a^{-1} \circ S \circ f_a$, is therefore a group isomorphism from $\mathrm{Aut}(\mathcal{M}, d_{\mathcal{M}}, P_{\mathcal{M}})$ onto $\mathrm{Aut}(\mathcal{Z}_a, d_a, P_a)$. Lemma B.6 applied to $T = A$ shows that the isomorphisms are the maps $A \circ f_a^{-1} \circ S \circ f_a$, and $A \circ f_a^{-1} = f_b^{-1}$ by Definition 2.1, which gives the first statement. The second follows because the latent automorphism group is trivial exactly when the data automorphism group is. □

Proposition B.8 (Density generically trivialises the automorphism group; Proposition 3.2)**.** *Let $(\mathcal{M}, g)$ be a compact connected Riemannian manifold, possibly with smooth boundary, and let G be its group of distance-preserving bijections, identified with $\mathrm{Iso}(\mathrm{int}\,\mathcal{M}, g)$ by Lemma A.11. The set of smooth positive densities for which the automorphism group is trivial,*

$$\mathcal{D}_{\mathrm{rigid}} := \{\rho \in C^\infty_{>0}(\mathcal{M}) \mid \mathrm{Aut}(\mathcal{M}, g, \rho\, \mathrm{dVol}_g) = \{\mathrm{id}\}\},$$

is open and dense in $C^\infty_{>0}(\mathcal{M})$ equipped with the Whitney C^∞ topology.

Proposition 3.2 is the case $g = g^{\mathcal{M}}$, with ρ the density of $P_{\mathcal{M}}$. By Lemma A.11, a density ρ lies in $\mathcal{D}_{\mathrm{rigid}}$ precisely when no non-trivial isometry $\varphi \in G \setminus \{\mathrm{id}\}$ leaves ρ invariant, i.e., there is no $\varphi \neq \mathrm{id}$ with $\rho \circ \varphi = \rho$ everywhere. The proof shows that the densities with a non-trivial stabiliser form a closed set with empty interior. Open and dense is the topological notion of genericity. No canonical measure exists on the infinite-dimensional space $C^\infty_{>0}(\mathcal{M})$, so we do not claim that almost every density satisfies C1.

Proof. All points, neighbourhoods and Lie derivatives below are taken in $\operatorname{int}\mathcal{M}$, on which G acts smoothly and effectively; this is possible because a non-trivial element of G moves some interior point, and invariance $\rho\circ\varphi=\rho$ on the dense set $\operatorname{int}\mathcal{M}$ is invariance on $\mathcal{M}$. By Lemma A.11, G is a compact Lie group with Lie algebra $\mathfrak{g}$[2] of dimension m and finite component group $\pi_0(G)=G/G_0$. Fix an inner product on $\mathfrak{g}$, with norm $\|\cdot\|$.

For each $\rho\in C^\infty_{>0}(\mathcal{M})$, define its *stabiliser* $\operatorname{Stab}(\rho):=\{\varphi\in G\mid\rho\circ\varphi=\rho\}$ as the closed Lie subgroup of G whose action leaves ρ invariant. Its Lie algebra is

$$K(\rho):=\operatorname{Lie}(\operatorname{Stab}(\rho))=\{Y\in\mathfrak{g}\mid L_{Y^\sharp}\rho=0\},$$

by Lemma A.36. Let $k(\rho):=\dim K(\rho)$, and let $T_\rho:\mathfrak{g}\to C^\infty(\mathcal{M})$ be the linear map $Y\mapsto L_{Y^\sharp}\rho$, so $k(\rho)=\dim\ker T_\rho$. Note that $\mathcal{D}_{\text{rigid}}=\{\rho\mid\operatorname{Stab}(\rho)=\{\mathrm{id}\}\}$.

Openness. We show $C^\infty_{>0}(\mathcal{M})\setminus\mathcal{D}_{\text{rigid}}$ is closed. Suppose $\rho_n\to\rho$ in C^∞ with each $\operatorname{Stab}(\rho_n)\neq\{\mathrm{id}\}$. Pick $\varphi_n\in\operatorname{Stab}(\rho_n)\setminus\{\mathrm{id}\}$. By compactness of G, we can pass to a subsequence (not relabelled) with $\varphi_n\to\varphi\in G$. Then $\rho_n\circ\varphi_n\to\rho\circ\varphi$ uniformly, since $|\rho_n(\varphi_n(p))-\rho(\varphi(p))|\le\|\rho_n-\rho\|_\infty+|\rho(\varphi_n(p))-\rho(\varphi(p))|$, where the first term tends to zero and the second does so uniformly in p because ρ is uniformly continuous and $\varphi_n\to\varphi$ uniformly. As $\rho_n\circ\varphi_n=\rho_n\to\rho$, this gives $\rho\circ\varphi=\rho$, so $\varphi\in\operatorname{Stab}(\rho)$.

If $\varphi\neq\mathrm{id}$, we are done. If $\varphi=\mathrm{id}$, then for large n, φ_n lies in a neighbourhood of id on which $\exp$ is a diffeomorphism; write $\varphi_n=\exp(t_nY_n)$ with $\|Y_n\|=1$, $t_n\to0^+$. Pass to a further subsequence with $Y_n\to Y$, $\|Y\|=1$. By the fundamental theorem of calculus applied to $s\mapsto\rho_n(\exp(sY_n)\cdot p)$ for any $p\in\mathcal{M}$,

$$0=\rho_n(\exp(t_nY_n)\cdot p)-\rho_n(p)=\int_0^{t_n}(L_{Y_n^\sharp}\rho_n)(\exp(sY_n)\cdot p)\,ds.$$

Dividing by t_n expresses 0 as the average of the integrand over $s\in[0,t_n]$. As $n\to\infty$, $\rho_n\to\rho$ in C^1, $Y_n\to Y$, and $\exp(sY_n)\cdot p\to p$ uniformly in $s\in[0,t_n]$; so the average tends to $(L_{Y^\sharp}\rho)(p)$. Hence $L_{Y^\sharp}\rho\equiv0$, so $Y\in K(\rho)$ with $\|Y\|=1$, giving $\exp(sY)\in\operatorname{Stab}(\rho)$ for all s, with $\exp(sY)\neq\mathrm{id}$ for small $s\neq0$ since $\exp$ is a local diffeomorphism at 0.

Density. We show every $\rho_0\in C^\infty_{>0}(\mathcal{M})$ lies in the closure of $\mathcal{D}_{\text{rigid}}$. The argument has two stages. The first reduces $k(\rho)$ to 0, so that $\operatorname{Stab}(\rho)$ becomes finite, and the second removes the remaining finite symmetries.

Step 1: Reduce k to 0. We prove the following lemma: *if $k(\rho)=\ell\ge1$, then every C^∞-neighbourhood of ρ contains some $\rho'\in C^\infty_{>0}(\mathcal{M})$ with $k(\rho')\le\ell-1$.* Iterating at most m times yields $\rho_1\in C^\infty_{>0}(\mathcal{M})$ arbitrarily close to ρ_0 with $k(\rho_1)=0$.

Let $V:=C^\infty(\mathcal{M})/\operatorname{Im}(T_\rho)$. For $\psi\in C^\infty(\mathcal{M})$, define

$$\overline{T}_\psi:K(\rho)\to V,\qquad Y\mapsto[L_{Y^\sharp}\psi].$$

Claim 1: for every nonzero $Y\in K(\rho)$ there exists $\psi\in C^\infty_c(\mathcal{M})$ with $\overline{T}_\psi(Y)\neq0$.

Since the action is effective and $Y\neq0$, there is $p\in\mathcal{M}$ with $Y^\sharp|_p\neq0$. Choose local coordinates $(x^1,\dots,x^d)$ on a neighbourhood U of p in which $Y^\sharp=\partial_{x^1}$ (straightening theorem). Consider the linear map

$$\Lambda:C^\infty_c(U)\to C^\infty(\mathcal{M}),\qquad\psi\mapsto L_{Y^\sharp}\psi=\partial_{x^1}\psi,$$

where functions on U are extended by zero to $\mathcal{M}$. The image $\operatorname{Im}(\Lambda)$ is infinite-dimensional. For instance, the functions $\partial_{x^1}(\chi\cdot(x^1)^j)=\chi\cdot j(x^1)^{j-1}+(\partial_{x^1}\chi)\cdot(x^1)^j$, for $j=1,2,3,\dots$ and a fixed bump $\chi\in C^\infty_c(U)$ with $\chi\equiv1$ near p, are linearly independent in $C^\infty(\mathcal{M})$ (their Taylor expansions at p have distinct leading orders).

On the other hand, $\operatorname{Im}(T_\rho)\subseteq C^\infty(\mathcal{M})$ is finite-dimensional, of dimension at most $m-\ell$. An infinite-dimensional subspace cannot be contained in a finite-dimensional subspace, so there exists $\psi\in C^\infty_c(U)$ with $L_{Y^\sharp}\psi\notin\operatorname{Im}(T_\rho)$, i.e., $\overline{T}_\psi(Y)\neq0$. This establishes Claim 1.

Fix a nonzero $Y_0\in K(\rho)$ (possible since $\ell\ge1$) and let ψ be as in Claim 1 for this Y_0. Fix any norm on the finite-dimensional subspace $\operatorname{Im}(T_\rho)+T_\psi(\mathfrak{g})\subset C^\infty(\mathcal{M})$, in which all vectors below live. For

[2]Here $\mathfrak{g}$ is the Lie algebra of G, not the Riemannian metric g.

$|s|$ small, $\rho + s\psi$ is positive, because $\rho > 0$ on the compact $\mathcal{M}$ and ψ is bounded, and $\rho + s\psi \to \rho$ in C^∞ as $s \to 0$.

Claim 2: $k(\rho + s\psi) \le \ell - 1$ for all sufficiently small $s \neq 0$.

Suppose not. Then there is a sequence $s_n \to 0$, $s_n \neq 0$, with $k(\rho + s_n\psi) \ge \ell$. Choose ℓ-dimensional subspaces $L_n \subseteq \ker T_{\rho+s_n\psi}$; by compactness of the Grassmannian of ℓ-planes in $\mathfrak{g}$, pass to a subsequence with $L_n \to L$, $\dim L = \ell$. We show $L \subseteq K(\rho)$ with $\overline{T}_\psi|_L \equiv 0$; since $\dim L = \ell = \dim K(\rho)$ this forces $L = K(\rho) \ni Y_0$, contradicting $\overline{T}_\psi(Y_0) \neq 0$.

Fix a unit $W \in L$ and choose unit $W_n \in L_n$ with $W_n \to W$. Decompose $W_n = W_n^{\parallel} + Z_n$ with $W_n^{\parallel} \in K(\rho)$, $Z_n \in K(\rho)^\perp$. Since $T_\rho|_{K(\rho)^\perp}$ is injective between finite-dimensional spaces, there is C with $\|Z\| \le C\|T_\rho(Z)\|$ on $K(\rho)^\perp$. From $0 = T_{\rho+s_n\psi}(W_n) = T_\rho(Z_n) + s_n T_\psi(W_n)$ we get $\|Z_n\| \le C|s_n|\,\|T_\psi(W_n)\| \le CM|s_n| \to 0$, where $M := \sup_{\|V\|=1}\|T_\psi(V)\|$; hence $W = \lim W_n^{\parallel} \in K(\rho)$. Moreover Z_n/s_n is bounded; pass to a further subsequence with $Z_n/s_n \to W' \in K(\rho)^\perp$. Dividing the displayed identity by s_n and letting $n \to \infty$ gives $T_\rho(W') = -T_\psi(W)$, so $T_\psi(W) \in \mathrm{Im}(T_\rho)$, i.e. $\overline{T}_\psi(W) = 0$. As $W \in L$ was an arbitrary unit vector, the claim follows.

Claims 1 and 2 prove the lemma, and iterating it completes Step 1. We obtain $\rho_1 \in C^\infty_{>0}(\mathcal{M})$ arbitrarily close to ρ_0 with $k(\rho_1) = 0$. Since $\mathrm{Stab}(\rho_1)$ is a closed subgroup of the compact group G with trivial Lie algebra, it is discrete, hence finite.

Step 2: Break the remaining finite symmetries. Let $\mathrm{Stab}(\rho_1) = \{\mathrm{id}, \varphi_1, \dots, \varphi_N\}$. Since $k(\rho_1) = 0$, the linear map T_{ρ_1} is injective, so $\mathrm{Im}(T_{\rho_1})$ is a subspace of dimension m.

Choice of the bumps. For each j, choose p_j with $\varphi_j(p_j) \neq p_j$, which exists because $\varphi_j \neq \mathrm{id}$ moves some interior point, and an open set $U_j \ni p_j$ with $U_j \cap \varphi_j(U_j) = \emptyset$, which exists because φ_j is continuous. A bump $\psi_j \in C_c^\infty(U_j)$ then sits at points that φ_j moves, but not at their images, as in Section 3. For bumps $\Psi = (\psi_1, \dots, \psi_N)$ with $\psi_i \in C_c^\infty(U_i)$, put

$$\Delta_j(\Psi) := \sum_{i=1}^{N} \big(\psi_i \circ \varphi_j - \psi_i\big), \qquad j = 1, \dots, N.$$

We choose Ψ with $\Delta_j(\Psi) \notin \mathrm{Im}(T_{\rho_1})$ for every j, which is possible for the following reason. Each Δ_j is linear in Ψ, so $B_j := \{\Psi : \Delta_j(\Psi) \in \mathrm{Im}(T_{\rho_1})\}$ is a linear subspace of $E := \prod_{i=1}^N C_c^\infty(U_i)$. It is a proper subspace. On the j-th factor alone, Δ_j is the map $\psi_j \mapsto \psi_j \circ \varphi_j - \psi_j$, and this map is injective on $C_c^\infty(U_j)$, because $\psi_j \circ \varphi_j$ vanishes on U_j when $U_j \cap \varphi_j(U_j) = \emptyset$, so $\psi_j \circ \varphi_j = \psi_j$ forces $\psi_j = 0$ on U_j and hence everywhere. Its image is therefore infinite-dimensional and not contained in the m-dimensional space $\mathrm{Im}(T_{\rho_1})$, so some element of E lies outside B_j. A real vector space is not a finite union of proper subspaces, so some $\Psi \in E$ lies outside $B_1 \cup \cdots \cup B_N$. Fix such a Ψ and write $\Delta_j := \Delta_j(\Psi)$.

For $\varepsilon > 0$ let

$$\rho_2^\varepsilon := \rho_1 + \varepsilon \sum_{j=1}^{N} \psi_j.$$

It is positive for small ε, because $\rho_1 > 0$ on the compact $\mathcal{M}$ and the ψ_j are bounded, and $\rho_2^\varepsilon \to \rho_1$ in C^∞ as $\varepsilon \to 0$. Since $\rho_1 \circ \varphi_j = \rho_1$,

$$\rho_2^\varepsilon \circ \varphi_j - \rho_2^\varepsilon = \varepsilon\,\Delta_j \qquad \text{for } j = 1, \dots, N, \tag{7}$$

and $\Delta_j \neq 0$ because $0 \in \mathrm{Im}(T_{\rho_1})$, so $\varphi_j \notin \mathrm{Stab}(\rho_2^\varepsilon)$ for every $\varepsilon > 0$. This is the bump that breaks φ_j. It remains to show that no new symmetry appears, that is, $\mathrm{Stab}(\rho_2^\varepsilon) = \{\mathrm{id}\}$ for all sufficiently small $\varepsilon > 0$. Then $\rho_2^\varepsilon \in \mathcal{D}_{\mathrm{rigid}}$ is C^∞-close to ρ_1, hence to ρ_0, which proves density.

No new symmetry appears. Suppose not. Then there are $\varepsilon_n \to 0$ and $\varphi_n \in \mathrm{Stab}(\rho_2^{\varepsilon_n}) \setminus \{\mathrm{id}\}$. Write $\rho_n := \rho_2^{\varepsilon_n}$, so $\rho_n \to \rho_1$ in C^∞. By compactness of G, pass to a subsequence with $\varphi_n \to \varphi \in G$. The first paragraph of the openness argument, applied to $\rho_n \to \rho_1$ and $\varphi_n \to \varphi$, gives $\varphi \in \mathrm{Stab}(\rho_1)$, so $\varphi = \mathrm{id}$ or $\varphi = \varphi_j$ for some j.

The limit is the identity. The second paragraph of the openness argument, applied to the same sequences, produces $Y \in K(\rho_1)$ with $\|Y\| = 1$. Then $k(\rho_1) \ge 1$, which contradicts $k(\rho_1) = 0$.

The limit is φ_j. Put $\chi_n := \varphi_j^{-1} \circ \varphi_n$, so $\chi_n \to \mathrm{id}$ and, for large n, $\chi_n = \exp(Y_n)$ with $Y_n \to 0$ in $\mathfrak{g}$. Since $\rho_1 \circ \varphi_j = \rho_1$, the identity $\rho_n \circ \varphi_n = \rho_n$ reads

$$\rho_1 \circ \chi_n - \rho_1 = \varepsilon_n \Big(\sum_i \psi_i - \sum_i \psi_i \circ \varphi_j \circ \chi_n \Big). \tag{8}$$

If $Y_n = 0$ for infinitely many n, then $\varphi_n = \varphi_j$ along them, and Equation 7 gives $\varepsilon_n \Delta_j = 0$, contradicting $\Delta_j \neq 0$. So $Y_n \neq 0$ for large n, and we write $Y_n = t_n \hat{Y}_n$ with $t_n := \|Y_n\| \to 0^+$ and $\|\hat{Y}_n\| = 1$. Pass to a subsequence along which $\hat{Y}_n \to \hat{Y}$ with $\|\hat{Y}\| = 1$ and $t_n/\varepsilon_n \to \lambda \in [0, \infty]$. Fix $p \in \operatorname{int} \mathcal{M}$ and evaluate Equation 8 at p. As in the openness argument, the fundamental theorem of calculus gives

$$\rho_1(\chi_n(p)) - \rho_1(p) = \int_0^{t_n} (L_{\hat{Y}_n^\sharp} \rho_1)(\exp(s\hat{Y}_n) \cdot p)\, ds = t_n \big[(L_{\hat{Y}^\sharp} \rho_1)(p) + o(1) \big],$$

because $(Y, q) \mapsto (L_{Y^\sharp} \rho_1)(q)$ is continuous and $\exp(s\hat{Y}_n) \cdot p \to p$ uniformly in $s \in [0, t_n]$. On the right-hand side, $\chi_n(p) \to p$ and the ψ_i and φ_j are continuous, so

$$\varepsilon_n \Big(\sum_i \psi_i(p) - \sum_i \psi_i(\varphi_j(\chi_n(p))) \Big) = \varepsilon_n \big[-\Delta_j(p) + o(1) \big].$$

Hence

$$t_n \big[(L_{\hat{Y}^\sharp} \rho_1)(p) + o(1) \big] = \varepsilon_n \big[-\Delta_j(p) + o(1) \big]. \tag{9}$$

If $\lambda = \infty$, divide Equation 9 by t_n and let $n \to \infty$. This gives $(L_{\hat{Y}^\sharp} \rho_1)(p) = 0$, and as p was arbitrary, $\hat{Y} \in K(\rho_1)$ with $\|\hat{Y}\| = 1$, which contradicts $k(\rho_1) = 0$. If $\lambda < \infty$, divide by ε_n instead. This gives $\lambda\, (L_{\hat{Y}^\sharp} \rho_1)(p) = -\Delta_j(p)$, and as p was arbitrary, $\Delta_j = T_{\rho_1}(-\lambda \hat{Y}) \in \operatorname{Im}(T_{\rho_1})$, which contradicts the choice of Ψ.

Both limits are impossible, so $\operatorname{Stab}(\rho_2^\varepsilon) = \{\mathrm{id}\}$ for all sufficiently small $\varepsilon > 0$. This completes the density argument. □

B.3 Proofs for Section 4

Proposition B.9 (Proposition 4.1)**.** *Under H1–H2, the optimal couplings of Equation 2 are exactly the couplings $(\mathrm{id} \times T)_\# P_a$ induced by isomorphisms T of $(\mathcal{Z}_a, d_a, P_a)$ and $(\mathcal{Z}_b, d_b, P_b)$. Under C1 the optimal coupling is unique and equals $\eta^\star = (\mathrm{id} \times A)_\# P_a$.*

Proof. The objective J_{GW} of Equation 2 is nonnegative.

Isomorphisms give optimal couplings. Let T be an isomorphism. Since $T_\# P_a = P_b$, the coupling $(\mathrm{id} \times T)_\# P_a$ lies in $\Pi(P_a, P_b)$, and since $d_b(T(z), T(z')) = d_a(z, z')$ for all z, z', its distortion vanishes. The alignment A is an isomorphism by Proposition 2.3, so the minimum is 0 and the optimal couplings are exactly those with $J_{\mathrm{GW}}(\eta) = 0$.

Optimal couplings come from isomorphisms. Let $\eta \in \Pi(P_a, P_b)$ with $J_{\mathrm{GW}}(\eta) = 0$, and let $S := \operatorname{supp}(\eta) \subseteq \mathcal{Z}_a \times \mathcal{Z}_b$. The integrand $|d_a(z, z') - d_b(w, w')|^2$ is continuous and nonnegative, and $\operatorname{supp}(\eta \otimes \eta) = S \times S$, so

$$d_a(z, z') = d_b(w, w') \qquad \text{for all } (z, w), (z', w') \in S. \tag{10}$$

The projection of the compact set S to $\mathcal{Z}_a$ is compact and carries full P_a-mass, so it contains $\operatorname{supp}(P_a) = \mathcal{Z}_a$, using that P_a has full support (Section 2). Likewise the projection to $\mathcal{Z}_b$ is $\mathcal{Z}_b$. If (z, w_1) and (z, w_2) lie in S, Equation 10 applied to these two pairs gives $d_b(w_1, w_2) = d_a(z, z) = 0$, so $w_1 = w_2$, and the same argument with the roles exchanged shows that $(z_1, w), (z_2, w) \in S$ forces $z_1 = z_2$. Hence S is the graph of a bijection $T \colon \mathcal{Z}_a \to \mathcal{Z}_b$, and Equation 10 says that T preserves distances, so T is continuous. A coupling with first marginal P_a that is supported on the graph of the continuous map T equals $(\mathrm{id} \times T)_\# P_a$, and its second marginal gives $T_\# P_a = P_b$. So T is an isomorphism and η is the coupling it induces.

Uniqueness under C1. By Proposition 3.1, C1 makes A the only isomorphism, so $\eta^\star$ is the only optimal coupling. □

Proposition B.10 (Proposition 4.2)**.** *Under H1–H2 and C2, $\eta^\star = (\mathrm{id} \times A)_\# P_a$ is the unique optimal coupling of the LDD-LP in Equation 3.*

Proof. Write $c(z, w) := W_2^2\big(\Phi_a(z, \cdot), \Phi_b(w, \cdot)\big)$ for the integrand of Equation 3, so that $J_{\mathrm{LDD}}(\eta) = \int c \,\mathrm{d}\eta$. Since A is an isomorphism (Proposition B.5), $\Phi_b(A(z), \cdot) = \Phi_a(z, \cdot)$ for every $z \in \mathcal{Z}_a$ (Section 4.1), so $\eta^\star \in \Pi(P_a, P_b)$ has cost $\int c \,\mathrm{d}\eta^\star = 0$. As $c \geq 0$, the minimum is zero and $\eta^\star$ attains it. Let $\eta \in \Pi(P_a, P_b)$ be any minimiser.

Under H1, distance spheres are $P_{\mathcal{M}}$-null (Remark A.16), so every LDD $\Phi(x, \cdot)$ is continuous in r, and $x \mapsto \Phi(x, \cdot)$ is continuous in the supremum norm (Lemma C.3). Hence the set $\{(x, x') : x \neq x',\ \Phi(x, \cdot) = \Phi(x', \cdot)\}$ is the intersection of a closed set with an open set in the compact space $\mathcal{M} \times \mathcal{M}$, hence σ-compact, and the collision set $\mathcal{C}$ of C2, its projection onto the first factor, is σ-compact and in particular Borel. Let $\mathcal{C}_a := \{z : \Phi_a(z', \cdot) = \Phi_a(z, \cdot) \text{ for some } z' \neq z\}$ be the collision set of $\mathcal{Z}_a$ and $\mathcal{C}_b$ that of $\mathcal{Z}_b$. Since f_a and f_b are isomorphisms onto the data mm-space, $\mathcal{C}_a = f_a^{-1}(\mathcal{C})$ and $\mathcal{C}_b = f_b^{-1}(\mathcal{C}) = A(\mathcal{C}_a)$, so $P_a(\mathcal{C}_a) = P_b(\mathcal{C}_b) = P_{\mathcal{M}}(\mathcal{C}) = 0$.

The cost c is continuous, because $W_2\big(\Phi_a(z, \cdot), \Phi_a(z', \cdot)\big) \leq d_a(z, z')$ by coupling the two distance distributions through one random data point, and likewise for b, and it is nonnegative, so $\int c \,\mathrm{d}\eta = 0$ forces $c(z, w) = 0$ for η-almost every (z, w). For such a pair, $\Phi_a(z, \cdot)$ and $\Phi_b(w, \cdot)$ agree for almost every r, and since both are continuous in r they agree for every r. Also $\eta(\mathcal{C}_a \times \mathcal{Z}_b) = P_a(\mathcal{C}_a) = 0$. Disintegrate $\eta = \int \delta_z \otimes \eta_z \,\mathrm{d}P_a(z)$. For P_a-almost every z, then, $z \notin \mathcal{C}_a$ and η_z is supported on $\{w : \Phi_b(w, \cdot) = \Phi_a(z, \cdot)\}$. This set contains $A(z)$ and no other point. A second $w \neq A(z)$ with $\Phi_b(w, \cdot) = \Phi_b(A(z), \cdot)$ would put $A(z)$ in $\mathcal{C}_b$, that is, $z \in \mathcal{C}_a$. Hence $\eta_z = \delta_{A(z)}$ for P_a-almost every z, and $\eta = (\mathrm{id} \times A)_\# P_a = \eta^\star$. □

The proof uses C2 directly and not C1. That C2 implies C1, as stated in Section 4.1, is Lemma C.1.

Lemma B.11 (Generic samples)**.** *Let $\mathrm{z}_1, \ldots, \mathrm{z}_n$, $n \geq 3$, be i.i.d. from P_a. Almost surely the sample points are distinct, the pairwise distances $d_a(\mathrm{z}_i, \mathrm{z}_j)$, $i < j$, are distinct, the rows of the distance matrix are distinct as multisets, and no non-trivial permutation of the sample preserves all pairwise distances.*

Proof. The distribution of representations $P_a = (f_a^{-1})_\# P_{\mathcal{M}}$ has a density with respect to the volume of g_a, because $P_{\mathcal{M}}$ has one with respect to the volume of $\mathcal{M}$ and f_a is an isometry onto $\mathcal{M}$. In particular P_a has no atoms, so the sample points are almost surely distinct. Distance spheres have volume zero (Remark A.16), so they have P_a-measure zero.

All $\binom{n}{2}$ pairwise distances are almost surely distinct. If two distinct index pairs satisfied $d_a(\mathrm{z}_p, \mathrm{z}_q) = d_a(\mathrm{z}_r, \mathrm{z}_s)$, some index would lie in exactly one of the pairs. Conditionally on all other sample points, that point is forced onto a distance sphere, which has probability zero. Integrating over the conditioning and taking a union over the finitely many pairs of index pairs gives the claim.

Let $R_i := \{d_a(\mathrm{z}_i, \mathrm{z}_l) : l \neq i\}$ be the multiset of entries of row i. For $i \neq j$, the multisets R_i and R_j share the entry $d_a(\mathrm{z}_i, \mathrm{z}_j)$, and their remaining entries are distances between distinct index pairs, hence pairwise distinct. Since $n \geq 3$, each row has at least one remaining entry, so $R_i \neq R_j$.

Finally, suppose a permutation $\tau \neq \mathrm{id}$ preserves all pairwise distances, and pick i with $\tau(i) \neq i$. Then $R_{\tau(i)} = \{d_a(\mathrm{z}_{\tau(i)}, \mathrm{z}_{\tau(l)}) : l \neq i\} = \{d_a(\mathrm{z}_i, \mathrm{z}_l) : l \neq i\} = R_i$, which contradicts the distinctness of rows. □

Proposition B.12 (Proposition 4.4)**.** *In the permutation regime, under H1–H2 and with $n \geq 3$, the empirical J_β almost surely has the unique optimal coupling $\eta_\sigma := \frac{1}{n}\sum_i \delta_{(\mathrm{z}_i^a, \mathrm{z}_{\sigma(i)}^b)}$ for every $\beta \in [0, 1]$. This holds whether or not C1 holds.*

Proof. Write $\eta \in \Pi(\hat{P}_a, \hat{P}_b)$ as an $n \times n$ matrix with row and column sums $1/n$, the empirical GW objective as $J_{\mathrm{GW}}(\eta) = \sum_{i,j,k,l} \left|C_{ik}^a - C_{jl}^b\right|^2 \eta_{ij}\eta_{kl}$, and the empirical LDD-LP cost as $c_{ij} = W_2^2\big(\hat{\Phi}_a(\mathrm{z}_i^a, \cdot), \hat{\Phi}_b(\mathrm{z}_j^b, \cdot)\big)$, where $\hat{\Phi}_a(\mathrm{z}_i^a, r) = \frac{1}{n}\#\{k : C_{ik}^a \leq r\}$ is the empirical distribution function of row i of C^a, and likewise for b. Both objectives are nonnegative. We work on the almost-sure event of Lemma B.11.

The coupling along σ is optimal for both problems. Since A is an isometry, $C^b_{\sigma(i)\sigma(k)} = d_b(A(\mathrm{z}^a_i), A(\mathrm{z}^a_k)) = d_a(\mathrm{z}^a_i, \mathrm{z}^a_k) = C^a_{ik}$, so $J_{\mathrm{GW}}(\eta_\sigma) = 0$. The same identity shows that row $\sigma(i)$ of C^b and row i of C^a are equal as multisets. Their empirical distribution functions therefore coincide, $c_{i\sigma(i)} = 0$, and η_σ has zero LDD cost.

GW has no other optimum. Let $J_{\mathrm{GW}}(\eta) = 0$, and let $S := \{(i,j) : \eta_{ij} > 0\}$. Every term of J_{GW} with $(i,j),(k,l) \in S$ vanishes, so $C^a_{ik} = C^b_{jl}$ for all such pairs. If (i,j) and (i,j') lie in S, this gives $C^b_{jj'} = C^a_{ii} = 0$, so $j = j'$ because the sample points are distinct, and symmetrically $(i,j),(i',j) \in S$ forces $i = i'$. The marginal constraints put every i and every j in at least one pair of S, so S is the graph of a permutation σ' with $C^b_{\sigma'(i)\sigma'(k)} = C^a_{ik}$. The permutation $\tau := \sigma^{-1} \circ \sigma'$ then satisfies $C^a_{\tau(i)\tau(k)} = C^b_{\sigma'(i)\sigma'(k)} = C^a_{ik}$, so it preserves all pairwise distances of the a-sample, and Lemma B.11 gives $\tau = \mathrm{id}$. Hence $\sigma' = \sigma$, and the marginals force $\eta = \eta_\sigma$.

The LDD-LP has no other optimum. Two distributions on the line are at W_2 distance zero only if they are equal, and an empirical distribution function on a finite multiset determines the multiset. So $c_{ij} = 0$ exactly when row i of C^a and row j of C^b are equal as multisets. Row j of C^b equals row $\sigma^{-1}(j)$ of C^a, and the rows of C^a are distinct, so $c_{ij} = 0$ exactly when $j = \sigma(i)$. A coupling with zero cost is therefore supported on the graph of σ, and the marginals force it to be η_σ.

The fused problem has no other optimum. The empirical objective $J_\beta(\eta) = (1-\beta)J_{\mathrm{GW}}(\eta) + \beta \sum_{i,j} c_{ij}\eta_{ij}$ is GW at $\beta = 0$ and the LDD-LP at $\beta = 1$. For $\beta \in (0,1)$ it is nonnegative and vanishes at η_σ by the first part, and its zeros are zeros of J_{GW}, so η_σ is its unique optimum by the second part.

None of the arguments uses C1 or C2. □

Lemma B.13. *In the permutation regime under H1–H2, almost surely for all sufficiently large n, the alignment A is the only distance-preserving bijection $T\colon (\mathcal{Z}_a, d_a) \to (\mathcal{Z}_b, d_b)$ with $T(\mathrm{z}^a_i) = \mathrm{z}^b_{\sigma(i)}$ for all i.*

Proof. Let T be such a bijection. Since $A(\mathrm{z}^a_i) = \mathrm{z}^b_{\sigma(i)}$, the map $\varphi := A^{-1} \circ T$ is a distance-preserving bijection of $(\mathcal{Z}_a, d_a)$ that fixes every z^a_i. By Lemma A.11 its restriction to $\operatorname{int} \mathcal{Z}_a$ is a Riemannian isometry. It remains to find sample points $x_1, \dots, x_{d+1}$ as in Corollary A.7. Fix $p \in \operatorname{int} \mathcal{Z}_a$. It has a totally normal neighbourhood $W \subset \operatorname{int} \mathcal{Z}_a$, so for some $\delta > 0$ and every $x \in W$, $\exp_x$ is a diffeomorphism from the δ-ball of $T_x\mathcal{Z}_a$ onto a set containing W [do Carmo, 1992, Chapter 3, Theorem 3.7]. Since P_a has full support, $P_a(W) > 0$, and by the strong law of large numbers almost surely at least $d+1$ of the z^a_i lie in W for all large n. Let x_1 be the first of them. The others lie in a normal ball around x_1, and given x_1 they are independent draws from P_a conditioned on W, which is absolutely continuous. By the remark following Corollary A.7, their log-map images at x_1 are then linearly independent almost surely. Corollary A.7 gives $\varphi = \mathrm{id}$, so $T = A$. □

Proposition B.14 (Consistency in the independent regime)**.** *In the independent regime under H1–H2, let $n \to \infty$.* (i) *Under C1, every sequence of global minimisers $\hat{\eta}_n$ of the empirical GW problem, or of the empirical J_β for a fixed $\beta \in (0,1)$, converges weakly to $\eta^\star$ almost surely.* (ii) *Under C2, the same holds for minimisers of the empirical LDD-LP.* (iii) *In each case, let $\hat{T}_n$ minimise $\int d_b\big(T(z), w\big)^2 \, \mathrm{d}\hat{\eta}_n(z,w)$ over a class $\mathcal{T}$ of continuous maps $\mathcal{Z}_a \to \mathcal{Z}_b$ that is compact in the uniform topology and contains A. Then $\hat{T}_n \to A$ uniformly on $\mathcal{Z}_a$ almost surely.*

Proof. All spaces are compact metric spaces, so weak convergence of probability measures on them is metrised by the Wasserstein distance W_1 [Villani, 2009, Theorem 6.9], and the space of probability measures on $\mathcal{Z}_a \times \mathcal{Z}_b$ is compact. The empirical measures $\hat{P}^n_a$ and $\hat{P}^n_b$ converge weakly to P_a and P_b almost surely [Billingsley, 1999]. We fix an outcome in this event and in the almost-sure events used below.

Approximating couplings. For every n there is $\eta_n \in \Pi(\hat{P}^n_a, \hat{P}^n_b)$ with $\eta_n \to \eta^\star$ weakly. To construct it, take optimal couplings $\alpha_n \in \Pi(\hat{P}^n_a, P_a)$ and $\beta_n \in \Pi(P_b, \hat{P}^n_b)$ for W_1, glue α_n, $\eta^\star$ and β_n along their shared marginals [Villani, 2009, Chapter 1], and let η_n be the marginal of the glued measure on the first and last factors. Then $\eta_n \in \Pi(\hat{P}^n_a, \hat{P}^n_b)$ and $W_1(\eta_n, \eta^\star) \le W_1(\hat{P}^n_a, P_a) + W_1(P_b, \hat{P}^n_b) \to 0$, for the sum metric on the product.

(i) The objective J_{GW} of Equation 2 is defined for every probability measure η on $\mathcal{Z}_a \times \mathcal{Z}_b$. Its integrand is continuous and bounded, and $\eta_n \to \eta$ weakly implies $\eta_n \otimes \eta_n \to \eta \otimes \eta$ weakly, so J_{GW} is weakly continuous. The empirical GW problem minimises J_{GW} over $\Pi(\hat{P}_a^n, \hat{P}_b^n)$, so $J_{\mathrm{GW}}(\hat{\eta}_n) \le J_{\mathrm{GW}}(\eta_n) \to J_{\mathrm{GW}}(\eta^\star) = 0$. Let $\bar{\eta}$ be the weak limit of a subsequence of $(\hat{\eta}_n)$, which exists by compactness. Its marginals are the limits of the empirical marginals, so $\bar{\eta} \in \Pi(P_a, P_b)$, and $J_{\mathrm{GW}}(\bar{\eta}) = 0$ by continuity. Under C1, Proposition 4.1 gives $\bar{\eta} = \eta^\star$. Every subsequence thus has a further subsequence converging to $\eta^\star$, so the whole sequence does.

(ii) Write $\Phi_a(z, r) := P_a(\bar{B}_a(z, r))$ and $\hat{\Phi}_a^n(z, r) := \hat{P}_a^n(\bar{B}_a(z, r))$, and likewise for b, for r in $[0, R]$ with R the diameter of $\mathcal{M}$, which $\mathcal{Z}_a$ and $\mathcal{Z}_b$ share. We first show $\sup_{z,r} |\hat{\Phi}_a^n(z, r) - \Phi_a(z, r)| \to 0$ almost surely. The map $z \mapsto \Phi_a(z, \cdot)$ is continuous in the supremum norm (Lemma C.3), so its image is compact in the continuous functions on $[0, R]$, hence equicontinuous, and Φ_a is uniformly continuous on $\mathcal{Z}_a \times [0, R]$. Given $\varepsilon > 0$, choose $\delta > 0$ with $|\Phi_a(z, r) - \Phi_a(z', r')| < \varepsilon$ whenever $d_a(z, z') \le \delta$ and $|r - r'| \le \delta$, and a finite δ-net $z_1, \dots, z_K$ of $\mathcal{Z}_a$. For $d_a(z, z_k) \le \delta$ the triangle inequality gives $\bar{B}_a(z_k, r - \delta) \subseteq \bar{B}_a(z, r) \subseteq \bar{B}_a(z_k, r + \delta)$, hence $\hat{\Phi}_a^n(z_k, r - \delta) \le \hat{\Phi}_a^n(z, r) \le \hat{\Phi}_a^n(z_k, r + \delta)$. For each of the finitely many z_k, $r \mapsto \hat{\Phi}_a^n(z_k, r)$ is the empirical distribution function of $d_a(z_k, \cdot)$ under $\hat{P}_a^n$, which converges uniformly in r almost surely by the Glivenko–Cantelli theorem. Combining the sandwich with uniform continuity gives $\limsup_n \sup_{z,r} |\hat{\Phi}_a^n(z, r) - \Phi_a(z, r)| \le \varepsilon$, and letting $\varepsilon \to 0$ along a sequence proves the claim. The empirical cost $c_n(z, w) := W_2^2\big(\hat{\Phi}_a^n(z, \cdot), \hat{\Phi}_b^n(w, \cdot)\big)$ therefore converges uniformly to the population cost c, because $W_2^2 \le R\, W_1 \le R^2 \|F - F'\|_\infty$ for two distributions on $[0, R]$ with distribution functions F and F', and $|W_2^2(\mu, \nu) - W_2^2(\mu', \nu')| \le 2R\big(W_2(\mu, \mu') + W_2(\nu, \nu')\big)$ for distributions on $[0, R]$. The population cost is continuous because $W_2\big(\Phi_a(z, \cdot), \Phi_a(z', \cdot)\big) \le d_a(z, z')$, by coupling the two distance distributions through one random data point. Then

$$\int c \,\mathrm{d}\hat{\eta}_n \le \int c_n \,\mathrm{d}\hat{\eta}_n + \|c_n - c\|_\infty \le \int c_n \,\mathrm{d}\eta_n + \|c_n - c\|_\infty \le \int c \,\mathrm{d}\eta_n + 2\|c_n - c\|_\infty \to \int c \,\mathrm{d}\eta^\star = 0.$$

Any weak limit point $\bar{\eta}$ of $(\hat{\eta}_n)$ lies in $\Pi(P_a, P_b)$ and has $\int c \,\mathrm{d}\bar{\eta} = 0$, so it is optimal for the LDD-LP, and under C2 Proposition 4.2 gives $\bar{\eta} = \eta^\star$. The subsequence argument of (i) concludes.

The fused problem. For $\beta \in (0, 1)$, let $\hat{\eta}_n$ minimise $J_\beta^n(\eta) := (1 - \beta) J_{\mathrm{GW}}(\eta) + \beta \int c_n \,\mathrm{d}\eta$ over $\Pi(\hat{P}_a^n, \hat{P}_b^n)$. Then $(1 - \beta) J_{\mathrm{GW}}(\hat{\eta}_n) \le J_\beta^n(\hat{\eta}_n) \le J_\beta^n(\eta_n) \to 0$, by the continuity of J_{GW} in (i) and the uniform convergence of c_n in (ii), so every limit point $\bar{\eta}$ has $J_{\mathrm{GW}}(\bar{\eta}) = 0$, and under C1 it equals $\eta^\star$ by Proposition 4.1. The subsequence argument of (i) concludes.

(iii) Let $F_\eta(T) := \int d_b(T(z), w)^2 \,\mathrm{d}\eta(z, w)$, and let D be the diameter of $\mathcal{Z}_b$. For maps T, T', $|F_\eta(T) - F_\eta(T')| \le 2D \sup_z d_b(T(z), T'(z))$ for every η, so each F_η is continuous on $\mathcal{T}$ and minimisers exist. Given $\varepsilon > 0$, take a finite ε-net $T_1, \dots, T_K$ of $\mathcal{T}$ in the uniform metric. Each $(z, w) \mapsto d_b(T_k(z), w)^2$ is continuous and bounded, so $F_{\hat{\eta}_n}(T_k) \to F_{\eta^\star}(T_k)$ by (i) or (ii), and

$$\sup_{T \in \mathcal{T}} \big|F_{\hat{\eta}_n}(T) - F_{\eta^\star}(T)\big| \le 4D\varepsilon + \max_k \big|F_{\hat{\eta}_n}(T_k) - F_{\eta^\star}(T_k)\big|.$$

Hence $F_{\hat{\eta}_n} \to F_{\eta^\star}$ uniformly on $\mathcal{T}$. The limit is $F_{\eta^\star}(T) = \int d_b(T(z), A(z))^2 \,\mathrm{d}P_a(z)$, which vanishes exactly when $T = A$ P_a-almost everywhere, and hence everywhere since T and A are continuous and P_a has full support. Let $\bar{T}$ be the uniform limit of a subsequence of $(\hat{T}_n)$, which exists by compactness. Then $F_{\eta^\star}(\bar{T}) = \lim F_{\hat{\eta}_{n_k}}(\hat{T}_{n_k}) \le \lim F_{\hat{\eta}_{n_k}}(A) = F_{\eta^\star}(A) = 0$, so $\bar{T} = A$, and the subsequence argument concludes. □

The next result makes the margin quantitative for the LDD-LP. It needs one hypothesis beyond H1–H2, a uniform bound on how fast ball masses grow,

$$\big|\Phi(x, r) - \Phi(x, r')\big| \le K\,|r - r'| \qquad \text{for all } x \in \mathcal{M} \text{ and } r, r' \ge 0, \tag{11}$$

that is, the distance from a point to a random data point has a density bounded by K, uniformly in the point. On a closed manifold this follows from H1. Let $\|\rho\|_\infty$ bound the density of $P_{\mathcal{M}}$ and let $S(x, s)$ be the distance sphere of radius s around x. For $r' < r$, the coarea formula of Appendix A, with $|\nabla d_{\mathcal{M}}(x, \cdot)| = 1$ almost everywhere, gives

$$\Phi(x, r) - \Phi(x, r') = P_{\mathcal{M}}\big(\{r' < d_{\mathcal{M}}(x, \cdot) \le r\}\big) \le \|\rho\|_\infty \int_{r'}^{r} \mathcal{H}^{d-1}\big(S(x, s)\big) \,\mathrm{d}s.$$

The sphere $S(x,s)$ is contained in $\exp_x(\{v \in T_x\mathcal{M} : |v| = s\})$, whose $(d-1)$-dimensional area is at most s^{d-1} times the integral over the unit sphere of $T_x\mathcal{M}$ of $\|\mathrm{d}(\exp_x)_{sv}\|^{d-1}$ by the area formula. This is bounded uniformly in x and $s \le R$, because exp is smooth on the compact set $\{w \in T\mathcal{M} : |w| \le R\}$, so Equation 11 holds with K equal to $\|\rho\|_\infty$ times that bound. With a boundary we state it as a hypothesis. Throughout, R is the diameter of $\mathcal{M}$, which is also the diameter of $\mathcal{Z}_a$ and $\mathcal{Z}_b$, LDDs are functions on $[0,R]$, the empirical LDDs are $\hat{\Phi}_a(z,r) := \hat{P}_a(\bar{B}_a(z,r))$ and likewise for b, c_n is the empirical LDD-LP cost built from them, and

$$\delta_n := \sup_z W_2\big(\hat{\Phi}_a(z,\cdot), \Phi_a(z,\cdot)\big) + \sup_w W_2\big(\hat{\Phi}_b(w,\cdot), \Phi_b(w,\cdot)\big)$$

is the estimation error of the LDDs, with W_2 between the distributions on $[0,R]$ that have these distribution functions. For such distributions $W_2^2 \le R\,W_1 \le R^2\|F - F'\|_\infty$, so δ_n is at most $R\sqrt{2}$ times the square root of the estimation error in the supremum norm over the point and the radius. The bounds below hold for any two samples. The sampling regime enters only through the rates at which $W_1(\hat{P}_a, P_a)$, $W_1(\hat{P}_b, P_b)$ and δ_n vanish, and in the independent regime all three tend to zero almost surely by the proof of Proposition B.14 and the inequality above.

Proposition B.15 (Stability of the LDD-LP)**.** *Assume H1–H2 and Equation 11 with constant K, and let R be the diameter of $\mathcal{M}$.* (i) *For every probability measure η on $\mathcal{Z}_a \times \mathcal{Z}_b$ and every t with $m(t) > 0$,*

$$\eta\big(\{(z,w) : d_b(w, A(z)) \ge t\}\big) \;\le\; \frac{K^2}{m(t)^3} \int c \,\mathrm{d}\eta. \tag{12}$$

(ii) *For any two samples, a minimiser $\hat{\eta}_n$ of the empirical LDD-LP satisfies*

$$\int c \,\mathrm{d}\hat{\eta}_n \;\le\; 2R\Big[W_1(\hat{P}_a, P_a) + W_1(\hat{P}_b, P_b) + 2\,\delta_n\Big], \tag{13}$$

with W_1 taken in d_a and d_b. (iii) *Let $B_n(t)$ be $K^2/m(t)^3$ times the right-hand side of Equation 13. In the metric $\max(d_a, d_b)$ on $\mathcal{Z}_a \times \mathcal{Z}_b$,*

$$d_P(\hat{\eta}_n, \eta^\star) \;\le\; \inf_{t>0} \max\big\{t,\; B_n(t)\big\} + d_P(\hat{P}_a, P_a).$$

Proof. Write $c(z,w) = W_2^2\big(\Phi_a(z,\cdot), \Phi_b(w,\cdot)\big)$ for the population cost and recall $\Phi_a(z,\cdot) = \Phi(f_a(z),\cdot)$ and $\Phi_b(w,\cdot) = \Phi(f_b(w),\cdot)$. For two distributions on the line, the squared Wasserstein distance is the squared L^2 distance between their quantile functions $F^{-1}(u) := \inf\{r : F(r) \ge u\}$ [Villani, 2009], as in Section 4.1.

LDDs are Lipschitz in the point. If $d_{\mathcal{M}}(x,x') = u$, the triangle inequality gives $\bar{B}_{\mathcal{M}}(x', r-u) \subseteq \bar{B}_{\mathcal{M}}(x,r) \subseteq \bar{B}_{\mathcal{M}}(x', r+u)$, and Equation 11 gives $\|\Phi(x,\cdot) - \Phi(x',\cdot)\|_\infty \le Ku$. Since f_a and f_b are isometries onto $\mathcal{M}$, the same holds for Φ_a in d_a and for Φ_b in d_b.

(i) Mismatched pairs have large cost. Let (z,w) satisfy $d_b(w, A(z)) \ge t$, and put $x := f_a(z) = f_b(A(z))$ and $y := f_b(w)$. Then $d_{\mathcal{M}}(x,y) = d_b(A(z), w) \ge t$, so $F := \Phi(x,\cdot)$ and $G := \Phi(y,\cdot)$ satisfy $\delta := \|F - G\|_\infty \ge m(t)$. By Equation 11, both are K-Lipschitz, so the supremum is attained at some r_0, and after exchanging the two we may assume $F(r_0) = G(r_0) + \delta$. Let $u \in (G(r_0), F(r_0))$, an interval of length δ. Since F is continuous, $F(F^{-1}(u)) = u$, and $F(r_0) \ge u$ gives $F^{-1}(u) \le r_0$; the Lipschitz bound $F(r_0) - u \le K\big(r_0 - F^{-1}(u)\big)$ then gives $F^{-1}(u) \le r_0 - (F(r_0) - u)/K$. In the same way, $G(r_0) < u$ gives $G^{-1}(u) > r_0$ and $u - G(r_0) \le K\big(G^{-1}(u) - r_0\big)$, so $G^{-1}(u) \ge r_0 + (u - G(r_0))/K$. Hence $G^{-1}(u) - F^{-1}(u) \ge \delta/K$ on an interval of length δ, and

$$c(z,w) = \int_0^1 \big(F^{-1}(u) - G^{-1}(u)\big)^2 \,\mathrm{d}u \ge \delta\Big(\frac{\delta}{K}\Big)^2 = \frac{\delta^3}{K^2} \ge \frac{m(t)^3}{K^2}.$$

Markov's inequality gives Equation 12 for every probability measure η.

The cost is Lipschitz, and close to its empirical version. Let $x, x' \in \mathcal{M}$ with $d_{\mathcal{M}}(x,x') = u$ and let $Y \sim P_{\mathcal{M}}$. The pair $\big(d_{\mathcal{M}}(x,Y), d_{\mathcal{M}}(x',Y)\big)$ couples the two distance distributions, and $|d_{\mathcal{M}}(x,Y) - d_{\mathcal{M}}(x',Y)| \le u$ by the triangle inequality, so $W_2\big(\Phi(x,\cdot), \Phi(x',\cdot)\big) \le u$. All distances lie in $[0,R]$, so every W_2 between two LDDs is at most R, and $|a^2 - b^2| \le 2R|a-b|$ for $a, b \in [0,R]$.

The triangle inequality for W_2 then gives $|c(z,w) - c(z',w')| \le 2R\big(d_a(z,z') + d_b(w,w')\big)$, and, with $c_n(z,w) = W_2^2\big(\hat\Phi_a(z,\cdot), \hat\Phi_b(w,\cdot)\big)$, $\varepsilon_n := \sup_{z,w} |c_n(z,w) - c(z,w)| \le 2R\,\delta_n$.

(ii) The computed coupling has small population cost. Let $\alpha \in \Pi(\hat P_a, P_a)$ and $\beta \in \Pi(\hat P_b, P_b)$ be W_1-optimal couplings in d_a and d_b. Gluing α, $\eta^\star$ and β along their common marginals gives a probability measure on quadruples $(\hat z, z, w, \hat w)$ with $(\hat z, z) \sim \alpha$, $(z,w) \sim \eta^\star$ and $(w, \hat w) \sim \beta$. Its $(\hat z, \hat w)$-marginal η_n lies in $\Pi(\hat P_a, \hat P_b)$, and the quadruple couples η_n with $\eta^\star$ at expected displacement $\mathbb{E}[d_a(\hat z, z) + d_b(\hat w, w)] = W_1(\hat P_a, P_a) + W_1(\hat P_b, P_b)$. Since $\int c\,\mathrm{d}\eta^\star = 0$, the Lipschitz bound on c gives $\int c\,\mathrm{d}\eta_n \le 2R\big(W_1(\hat P_a, P_a) + W_1(\hat P_b, P_b)\big)$. As $\hat\eta_n$ minimises $\int c_n$ over $\Pi(\hat P_a, \hat P_b)$ and $|c - c_n| \le \varepsilon_n$,

$$\int c\,\mathrm{d}\hat\eta_n \le \int c_n\,\mathrm{d}\hat\eta_n + \varepsilon_n \le \int c_n\,\mathrm{d}\eta_n + \varepsilon_n \le \int c\,\mathrm{d}\eta_n + 2\varepsilon_n \le 2R\Big[W_1(\hat P_a, P_a) + W_1(\hat P_b, P_b) + 2\,\delta_n\Big],$$

which is Equation 13.

(iii) Prokhorov distance. Let $S_t := \{(z,w) : d_b(w, A(z)) \ge t\}$, so that $\hat\eta_n(S_t) \le B_n(t)$ by (i) and (ii), and let $\tau(z,w) := (z, A(z))$, so that $\tau_\# \hat\eta_n = (\mathrm{id} \times A)_\# \hat P_a$. Off S_t, the point (z,w) lies at distance less than t from $\tau(z,w)$. For a Borel set B and its open s-neighbourhood B^s, with $s := \max\{t, B_n(t)\}$,

$$\hat\eta_n(B) \le \hat\eta_n(B \setminus S_t) + \hat\eta_n(S_t) \le (\tau_\#\hat\eta_n)(B^s) + s,$$

so $d_P\big(\hat\eta_n, (\mathrm{id} \times A)_\# \hat P_a\big) \le s$ [Billingsley, 1999]. The map $\mathrm{id} \times A$ is an isometric embedding of $(\mathcal{Z}_a, d_a)$ into $(\mathcal{Z}_a \times \mathcal{Z}_b, \max(d_a, d_b))$, because A is an isometry, so $d_P\big((\mathrm{id} \times A)_\# \hat P_a, \eta^\star\big) \le d_P(\hat P_a, P_a)$. The triangle inequality and the infimum over t conclude. □

The bound does not involve A on its right-hand side, and its margin is the supremum-norm margin of Definition 4.3 that the test of Section 4.1 estimates. If in addition $m(t) \ge a\,t^\alpha$ for small t, choosing t to balance the two terms of the Prokhorov bound turns it into a rate in the sampling and estimation errors. Of these, the W_1 terms decay as $n^{-1/d}$ for $d \ge 3$, the rate of Fournier and Guillin [2015] for measures on $\mathbb{R}^d$, which transfers to $(\mathcal{Z}_a, d_a)$ and $(\mathcal{Z}_b, d_b)$ through finitely many bi-Lipschitz charts of the compact d-dimensional $\mathcal{M}$. We do not evaluate either bound, since K, a and α are unknown for learned representations, and at practical sample sizes the constants make the bound uninformative. We give no analogous bound for GW, since C1 carries no margin.

B.3.1 Fused GW

This subsection proves the uniqueness of the optimum of the fused objective J_β of Equation 4 stated in Section 4.1, for $\beta \in [0,1)$ and with c the cost of Equation 3. For a coupling η, let $L_\eta(z,w) := \iint |d_a(z,z') - d_b(w,w')|^2\,\mathrm{d}\eta(z',w')$, so that the GW objective at η is $\langle L_\eta, \eta\rangle$ and the derivative of J_β at η in the direction $\eta' - \eta$ is $\langle 2(1-\beta)L_\eta + \beta c,\ \eta' - \eta\rangle$.

Lemma B.16 (The LDD-LP term is a pointwise lower bound)**.** *For every $\eta \in \Pi(P_a, P_b)$ and all (z,w), $c(z,w) \le L_\eta(z,w)$. In particular, the LDD-LP objective is at most the GW objective at every coupling.*

Proof. Pushing η through $(z',w') \mapsto \big(d_a(z,z'), d_b(w,w')\big)$ gives a coupling of the distance distributions of z and w, because the marginals of η are P_a and P_b. Its cost is $L_\eta(z,w)$, and $c(z,w) = W_2^2$ is the least cost of such a coupling. □

The lemma puts the two terms of J_β in the same units, so the weight β compares like with like and neither term needs rescaling.

Proposition B.17 (Unique optimum)**.** *Under H1–H2 and C1, $\eta^\star$ is the unique minimiser of J_β for every $\beta \in [0,1)$.*

Proof. Both terms are nonnegative. Both vanish at $\eta^\star$, the GW term because A is an isometry and the LDD-LP term because A preserves LDDs (Section 4.1). Since $1 - \beta > 0$, a zero of J_β is a zero of the GW objective, and under C1 the only one is $\eta^\star$ (Proposition 4.1). □

J_β is not convex for $\beta < 1$, since its second derivative is $(1-\beta)$ times that of the GW objective, which is negative in some directions, for example when two points exchange the points they are coupled to. We prove no guarantee for its stationary points. What the linear term adds is visible at the independent coupling $P_a \otimes P_b$, where solvers usually start.

Proposition B.18 (Linearisation at the independent coupling)**.** *Let $e_a(z)$ and $v_a(z)$ be the mean and variance of $d_a(z, Z')$ for $Z' \sim P_a$, and $e_b(w)$ and $v_b(w)$ those of $d_b(w, W')$ for $W' \sim P_b$. Then $L_{P_a \otimes P_b}(z, w) = \big(e_a(z) - e_b(w)\big)^2 + v_a(z) + v_b(w)$.*

Proof. Under $P_a \otimes P_b$, the distances $s = d_a(z, Z')$ and $s' = d_b(w, W')$ are independent, so $\mathbb{E}(s - s')^2 = \mathbb{E}s^2 + \mathbb{E}s'^2 - 2\,\mathbb{E}s\,\mathbb{E}s' = \big(e_a(z) - e_b(w)\big)^2 + v_a(z) + v_b(w)$. □

The terms $v_a(z)$ and $v_b(w)$ add the same constant to the cost of every coupling, so a transport step from $P_a \otimes P_b$ with this cost compares points only by their mean distance to all others. The step for J_β adds βc to the cost and compares their whole LDDs.

B.4 Symmetry breaking in vMF mixtures

Section 3 builds P_{mix} from three vMF components whose placement breaks every symmetry of the sphere (Figure 2, right). This subsection proves that it does, for vMF mixtures on S^d. For $c \in S^d \subset \mathbb{R}^{d+1}$ and $\kappa > 0$, the vMF density with respect to the round volume is $p_{c,\kappa}(x) = C_d(\kappa)\, e^{\kappa \langle c, x \rangle}$, with $C_d(\kappa)$ the normalising constant. Every isometry of the round sphere is the restriction of an orthogonal map $Q \in O(d+1)$ [do Carmo, 1992]. Q preserves the volume, and $\langle c, Q^{-1}x \rangle = \langle Qc, x \rangle$, so it maps one component onto another, $Q_\#(p_{c,\kappa}\,\mathrm{dVol}) = p_{Qc,\kappa}\,\mathrm{dVol}$. A symmetry of a mixture therefore permutes its components, because distinct components are linearly independent:

Lemma B.19 (Linear independence of spherical exponentials)**.** *Let $\theta_1, \dots, \theta_K \in \mathbb{R}^{d+1}$ ($d \geq 1$) be pairwise distinct. Then the functions $x \mapsto e^{\langle \theta_m, x \rangle}$, $m = 1, \dots, K$, are linearly independent on S^d. Consequently, vMF components with pairwise distinct parameters (c_m, κ_m) are linearly independent, since $(c, \kappa) \mapsto \kappa c$ is injective on $S^d \times (0, \infty)$.*

Proof. Suppose $\sum_m a_m e^{\langle \theta_m, x \rangle} = 0$ for all $x \in S^d$. Choose $u \in S^d$ avoiding the finitely many proper hyperplanes $\{u : \langle \theta_m - \theta_l, u \rangle = 0\}$, $m \neq l$, so that the values $\alpha_m := \langle \theta_m, u \rangle$ are pairwise distinct, and fix any unit $v \perp u$, setting $\beta_m := \langle \theta_m, v \rangle$. Restricting to the great circle $\gamma(t) = \cos t \cdot u + \sin t \cdot v$ gives

$$h(t) := \sum_{m=1}^{K} a_m\, e^{\alpha_m \cos t + \beta_m \sin t} = 0 \qquad \forall\, t \in \mathbb{R}.$$

The function h extends to an entire function of $t \in \mathbb{C}$ and vanishes on $\mathbb{R}$, hence vanishes identically on $\mathbb{C}$. Evaluate at $t = -iy$ with $y > 0$. Since $\cos(-iy) = \cosh y$ and $\sin(-iy) = -i \sinh y$, the m-th term has modulus $|a_m|\, e^{\alpha_m \cosh y}$. Let $m^\star$ be the index of the (unique) largest α_m among those with $a_m \neq 0$; if any coefficient is nonzero, then dividing $h(-iy) = 0$ by $e^{\alpha_{m^\star} \cosh y}$ and letting $y \to \infty$ leaves $|a_{m^\star}|$ on one side and 0 on the other, a contradiction. Hence all $a_m = 0$. □

Proposition B.20 (Rigidity of vMF mixtures)**.** *Let $d \geq 2$ and let $\mu = \sum_{m=1}^{K} w_m\, p_{c_m,\kappa_m}\,\mathrm{dVol}$ on (S^d, g_{S^d}) with $K \geq d+1$, parameters (c_m, κ_m) pairwise distinct, $w_m > 0$, $\sum_m w_m = 1$. Assume*

- (i) *some $d+1$ of the centres $c_1, \dots, c_K$ are linearly independent in $\mathbb{R}^{d+1}$; and*
- (ii) *at least one of:* (a) *the weights w_m are pairwise distinct;* (b) *the concentrations κ_m are pairwise distinct;* (c) *the pairwise geodesic distances $d(c_i, c_j)$, $i < j$, are pairwise distinct.*

Then $\mathrm{Aut}(S^d, g_{S^d}, \mu) = \{\mathrm{id}\}$, so C1 holds.

Proof. Let $Q \in \mathrm{Aut}(S^d, g_{S^d}, \mu) \subseteq O(d+1)$, so that $\sum_m w_m\, p_{Qc_m,\kappa_m} = \sum_m w_m\, p_{c_m,\kappa_m}$. *Step 0: Q permutes the components.* The parameters (Qc_m, κ_m) are pairwise distinct, as are the (c_m, κ_m). Collect both sides over the union T of the two parameter sets, $\sum_{\tau \in T} a_\tau\, p_\tau = 0$, where a_τ is the weight of τ on the left minus its weight on the right, each at most one w_m. By Lemma B.19, every $a_\tau = 0$. For $\tau = (Qc_m, \kappa_m)$ the left contributes $w_m > 0$, so some unique $\pi(m)$ has

$(c_{\pi(m)}, \kappa_{\pi(m)}) = (Qc_m, \kappa_m)$ and $w_{\pi(m)} = w_m$. The map π is injective, since Q is, so it is a permutation.

Step 1: $\pi = \mathrm{id}$. Under (ii)(a), π preserves the pairwise distinct weights, so $\pi = \mathrm{id}$; under (ii)(b), the same argument applies to the concentrations. Under (ii)(c), since Q is an isometry, $d(c_{\pi(i)}, c_{\pi(j)}) = d(Qc_i, Qc_j) = d(c_i, c_j)$ for all $i < j$, and as all pairwise distances are distinct, π maps each unordered pair $\{i, j\}$ to itself. If $\pi \neq \mathrm{id}$, pick i with $\pi(i) = j \neq i$ and any $l \notin \{i, j\}$, which exists since $K \geq d + 1 \geq 3$. Pair preservation gives $\pi(i) \in \{i, l\}$, yet $\pi(i) = j \notin \{i, l\}$, a contradiction. Hence $\pi = \mathrm{id}$.

Step 2: $Q = I$. From $\pi = \mathrm{id}$ we get $Qc_m = c_m$ for all m. By (i), Q fixes $d + 1$ linearly independent vectors in $\mathbb{R}^{d+1}$, that is, a basis, so $Q = I$. □

Condition (i) cannot be dropped, since the reflection in a hyperplane that contains every centre fixes each of them and preserves the mixture. P_{mix} satisfies (i) and all three branches of (ii) (Appendix E.1), so C1 holds for it.

C The gap between C1 and C2

The two problems of Section 4.1 need different conditions on the data mm-space. GW compares all pairwise distances, and under symmetry breaking (C1) its optimal coupling is unique (Proposition 4.1). The LDD-LP compares points only through their LDDs, and its optimal coupling is unique under distinguishable points (C2) (Proposition 4.2). The second condition is the stronger one. A symmetry preserves distances and $P_{\mathcal{M}}$, so it preserves every LDD, and under C2 it cannot move any point:

Lemma C.1. *Under H1, C2 implies C1.*

Proof. Let S be a symmetry. It maps every closed ball onto the closed ball of the same radius around the image of the centre, and it preserves $P_{\mathcal{M}}$, so $\Phi(S(x), \cdot) = \Phi(x, \cdot)$ for every x. If $S(x) \neq x$, then x shares its LDD with the distinct point $S(x)$, so $x \in \mathcal{C}$. Hence $S = \mathrm{id}$ on $\mathcal{M} \setminus \mathcal{C}$, which has full measure by C2 and is therefore dense, since the positive density of H1 gives every open set positive mass. S is continuous, so $S = \mathrm{id}$ everywhere. □

C.1 Collisions without symmetries

The converse fails, because two points can share an LDD although no symmetry exchanges them. Small balls show how much such a collision requires. For x in the interior of $\mathcal{M}$ and small r, the mass of a ball expands as

$$\Phi(x, r) = \omega_d r^d \rho(x) \big[1 + c_2(x)\, r^2 + O(r^4)\big], \qquad c_2(x) = -\frac{\mathrm{Scal}(x)}{6(d+2)} + \frac{\Delta\rho(x)}{2(d+2)\,\rho(x)},$$

where $d = \dim \mathcal{M}$, ω_d is the volume of the unit ball of $\mathbb{R}^d$, ρ is the density of $P_{\mathcal{M}}$, and Scal and Δ are the scalar curvature and the Laplace–Beltrami operator of $g^{\mathcal{M}}$. Gray [1974] derives the expansion for $\rho \equiv 1$, and expanding ρ in normal coordinates around x adds the term with $\Delta\rho$. Higher orders add coefficients $c_4, c_6, \ldots$ built from the curvature, ρ and their derivatives at x, and each of them can be read off $\Phi(x, \cdot)$ as $r \to 0$. Two points with the same LDD therefore agree in ρ, in every coefficient and in the mass of every larger ball. These are infinitely many equations on the $2d$ coordinates of the pair, so for a density without special structure we expect collisions to be rare and C2 to hold. Unlike for C1 (Proposition 3.2), we do not prove this.

A map that preserves every LDD, however, produces a collision at every point it moves. A symmetry is such a map, and C1 rules it out. A map that preserves $P_{\mathcal{M}}$ and every LDD, but not distances, is another, and on the following torus it moves almost every point:

Proposition C.2. *Let $\mathcal{M}$ be the flat torus $\mathbb{R}^2/2\pi\mathbb{Z}^2$, embedded in $\mathbb{R}^4$ by $x \mapsto (\cos x_1, \sin x_1, \cos x_2, \sin x_2)$, and let $P_{\mathcal{M}}$ have the density*

$$\rho(x) = \frac{1}{4\pi^2}\Big(1 + \tfrac{1}{4}\cos x_1 + \tfrac{1}{4}\cos(x_1 + x_2) + \tfrac{1}{4}\sin 2x_1\Big)$$

with respect to its area. This data mm-space satisfies H1 and C1. The map $\tau(x) = (x_1, -2x_1 - x_2)$ satisfies $\Phi(\tau(x), \cdot) = \Phi(x, \cdot)$ for every x, and $\tau(x) \neq x$ unless $x_1 + x_2 \in \pi\mathbb{Z}$, so the collision set has full measure and C2 fails.

Proof. The embedding has orthonormal partial derivatives, so it induces the flat metric $\mathrm{d}x_1^2 + \mathrm{d}x_2^2$, its area is the Lebesgue measure on $[0, 2\pi)^2$, and $d_{\mathcal{M}}$ is the distance of the flat torus. $\mathcal{M}$ is a smooth, compact and connected surface in $\mathbb{R}^4$. The density ρ is smooth, positive because the three trigonometric terms add up to at most $\frac{3}{4}$ in absolute value, and integrates to one because each of them integrates to zero, so H1 holds.

LDDs. Translations of the torus are isometries, so $d_{\mathcal{M}}(x, y) = \|y - x\|_{\mathcal{M}}$ with $\|v\|_{\mathcal{M}} := d_{\mathcal{M}}(0, v)$, and $\| - v\|_{\mathcal{M}} = \|v\|_{\mathcal{M}}$. Hence $\Phi(x, r) = \int \rho(x + v)\, \chi_r(v)\, \mathrm{d}v$ with $\chi_r(v) := \mathbf{1}\{\|v\|_{\mathcal{M}} \leq r\}$, which is even in v. For $k \in \mathbb{Z}^2$, evenness gives $\int \sin(k \cdot v)\, \chi_r(v)\, \mathrm{d}v = 0$, so with $\hat{\chi}_r(k) := \int \cos(k \cdot v)\, \chi_r(v)\, \mathrm{d}v$ the addition formulas give

$$\Phi(x, r) = \frac{1}{4\pi^2}\Big(\hat{\chi}_r(0) + \tfrac{1}{4}\hat{\chi}_r(1, 0)\cos x_1 + \tfrac{1}{4}\hat{\chi}_r(1, 1)\cos(x_1 + x_2) + \tfrac{1}{4}\hat{\chi}_r(2, 0)\sin 2x_1\Big).$$

The map τ is $x \mapsto Bx$ with $B = \left(\begin{smallmatrix} 1 & 0 \\ -2 & -1 \end{smallmatrix}\right)$. B has integer entries and determinant -1, so τ is a well-defined bijection of the torus, and $\tau \circ \tau = \mathrm{id}$. It leaves x_1 and $2x_1$ unchanged and maps $x_1 + x_2$ to $-(x_1 + x_2)$, so the formula above gives $\Phi(\tau(x), r) = \Phi(x, r)$ for every r. On the torus, $\tau(x) = x$ if and only if $2(x_1 + x_2) \in 2\pi\mathbb{Z}$, which holds on two closed curves, a null set. Every other point shares its LDD with the distinct point $\tau(x)$, so $\mathcal{C}$ has full measure.

Symmetry breaking. Let S be a symmetry. Since $\mathbb{R}^2$ is simply connected, S lifts through the quotient map $\mathbb{R}^2 \to \mathcal{M}$ to a continuous map $\tilde{S}$ of $\mathbb{R}^2$. The quotient map preserves distances between points less than π apart, so $\tilde{S}$ preserves Euclidean distances on a neighbourhood of every point and agrees there with a rigid motion. Two rigid motions that agree on an open set are equal, and $\mathbb{R}^2$ is connected, so $\tilde{S}(x) = Ox + b$ with $O \in O(2)$ and $b \in \mathbb{R}^2$. $\tilde{S}(x + 2\pi n)$ and $\tilde{S}(x)$ project to the same point for every $n \in \mathbb{Z}^2$, so $On \in \mathbb{Z}^2$, the columns of O are integer unit vectors, and O is a signed permutation matrix. S preserves the area, so $S_{\#}P_{\mathcal{M}} = P_{\mathcal{M}}$ and the continuity of ρ give $\rho \circ S = \rho$. Each term $\cos(k \cdot x)$ or $\sin(k \cdot x)$ of ρ becomes a term of frequency $O^\top k$ in $\rho \circ S$, since $k \cdot (Ox + b) = (O^\top k) \cdot x + k \cdot b$. The functions $e^{\mathrm{i}k \cdot x}$, $k \in \mathbb{Z}^2$, are orthogonal, so $\rho \circ S = \rho$ forces $O^\top$ to map the frequencies $\pm(1, 0)$, $\pm(1, 1)$, $\pm(2, 0)$ of ρ onto themselves. Their lengths 1, $\sqrt{2}$ and 2 differ, and $O^\top$ preserves lengths, so $O^\top(1, 0) = \pm(1, 0)$ and $O^\top(1, 1) = \pm(1, 1)$, which leaves $O = \pm I$. If $O = I$, the terms of frequency $(1, 0)$ in $\rho(x + b) = \rho(x)$ give $\cos(x_1 + b_1) = \cos x_1$, so $b_1 \in 2\pi\mathbb{Z}$, and those of frequency $(1, 1)$ give $b_1 + b_2 \in 2\pi\mathbb{Z}$, so $S = \mathrm{id}$. If $O = -I$, the terms of frequency $(1, 0)$ in $\rho(b - x) = \rho(x)$ give $\cos(b_1 - x_1) = \cos x_1$, so $b_1 \in 2\pi\mathbb{Z}$. The terms of frequency $(2, 0)$ then require $\sin(2b_1 - 2x_1)$, which equals $-\sin 2x_1$, to equal $\sin 2x_1$ for every x_1, which fails. Hence $S = \mathrm{id}$ and C1 holds. □

The map τ preserves the area, since $|\det B| = 1$, and $\rho \circ \tau = \rho$, so it preserves $P_{\mathcal{M}}$ and every LDD. It does not preserve distances, since for small $s > 0$ it maps the point $(s, 0)$, at distance s from $0 = \tau(0)$, to $(s, -2s)$, at distance $\sqrt{5}\, s$.

The sphere example of Section 3 also has collisions without symmetries. There, three vMF components with distinct weights and concentrations break every symmetry of the sphere, which gives P_{mix} (Figure 2, right). The experiments of Section 6 add the same mixture with equal weights and equal concentrations. It still satisfies C1, because its centres c_1, c_2, c_3 are linearly independent with distinct pairwise angles (Proposition B.20, (i) and (ii)(c)), but its LDDs collide on curves. The mass of a cap under one component depends only on the angle between the centre of the cap and that of the component. With equal weights and concentrations, $\Phi(x, \cdot)$ therefore depends only on the multiset of inner products $\{\langle x, c_m\rangle\}_{m=1}^3$. Let V be the matrix with rows c_1, c_2, c_3, and let $P \neq I$ be a permutation matrix. The point $y = V^{-1}PVx$ has the inner products of x in another order, so $\Phi(y, \cdot) = \Phi(x, \cdot)$ whenever $\|y\| = 1$. These x form the zero set on S^2 of the quadratic form $\|V^{-1}PVx\|^2 - \|x\|^2$. The form is not zero, because an orthogonal $V^{-1}PV$ would have a transpose that permutes the centres and preserves their pairwise angles, which their distinct angles rule out for $P \neq I$. Its zero set on S^2 is therefore null (Lemma A.17). Three of the colliding pairs are antipodal. For $\{i, j, k\} = \{1, 2, 3\}$, the point $x_k \propto (c_i + c_j) \times c_k$ has inner products a, $-a$ and 0 with c_i, c_j and c_k, with $a \neq 0$ because the centres are linearly independent, and $-x_k$ has the same multiset.

C.2 The margin

Since LDDs are estimated from samples, Section 4.1 requires them to be well separated and measures their separation by the margin $m(t)$ of Definition 4.3. The margin fixes a separation t, because nearby

points have nearby LDDs, so over all pairs of distinct points the smallest distance between LDDs is zero, with or without collisions:

Lemma C.3. *Under H1, the map $x \mapsto \Phi(x, \cdot)$ is continuous in the supremum norm.*

Proof. Under H1, distance spheres are $P_{\mathcal{M}}$-null (Remark A.16), so each $\Phi(x, \cdot)$ is continuous, and it equals 1 from the diameter R of $\mathcal{M}$ on. Fix x and x' and let $u = d_{\mathcal{M}}(x, x')$. By the triangle inequality, the closed balls of radius r around x and x' differ only within the shell $\{z : |d_{\mathcal{M}}(x, z) - r| \le u\}$, so $|\Phi(x, r) - \Phi(x', r)|$ is at most the mass of this shell. As $u \to 0$, the shells decrease to a distance sphere, which is null, so $\Phi(x', r) \to \Phi(x, r)$ for every r. For uniformity, let $\epsilon > 0$ and choose radii $0 = r_0 < \cdots < r_L = R$ with $\Phi(x, r_{l+1}) - \Phi(x, r_l) \le \epsilon$. For x' close enough to x, $|\Phi(x', r_l) - \Phi(x, r_l)| \le \epsilon$ at every r_l. For $r \in [r_l, r_{l+1}]$, monotonicity then gives $\Phi(x', r) \le \Phi(x', r_{l+1}) \le \Phi(x, r) + 2\epsilon$ and likewise $\Phi(x', r) \ge \Phi(x, r) - 2\epsilon$, so $\|\Phi(x', \cdot) - \Phi(x, \cdot)\|_\infty \le 2\epsilon$. □

At a fixed separation, a positive margin rules out every collision, and by compactness the converse holds as well:

Proposition C.4. *Under H1, the collision set $\mathcal{C}$ is empty if and only if $m(t) > 0$ for every $t \in (0, R]$.*

Proof. If $m(t) > 0$ for every t, two points $x \neq y$ are admissible in the infimum of $m(d_{\mathcal{M}}(x, y))$, so their LDDs differ. Conversely, let $\mathcal{C}$ be empty and suppose $m(t) = 0$ for some t. Choose pairs (x_n, y_n) with $d_{\mathcal{M}}(x_n, y_n) \ge t$ and $\|\Phi(x_n, \cdot) - \Phi(y_n, \cdot)\|_\infty \to 0$. By compactness, a subsequence converges to a pair (x, y) with $d_{\mathcal{M}}(x, y) \ge t$, so $x \neq y$, and by Lemma C.3 $\Phi(x, \cdot) = \Phi(y, \cdot)$, which contradicts $\mathcal{C} = \emptyset$. □

A positive margin at every separation is therefore stronger than C2, which allows a null collision set. One pair of colliding points at separation s makes $m(t) = 0$ for every $t \le s$. In the sphere example, the antipodal pairs of the equal-parameter mixture make its margin zero at every separation, and so do the colliding points of P_{unif} and P_{vMF}, which lie up to π apart.

The separation is also the resolution at which A is recovered. If the estimation error of the LDDs is small against $m(t)$, the LDD-LP puts little mass on pairs (z, w) with w at least t from $A(z)$ (Proposition B.15). Since m is non-decreasing in t, a finer resolution comes with a smaller margin.

C.3 The margin test

From samples we cannot compute the margin, but we can test whether two query points at least t apart share an LDD. The test compares the smallest distance between their empirical LDDs with the accuracy that the DKW inequality guarantees for each of them:

Lemma C.5. *Let $s_1, \ldots, s_N$ and $q_1, \ldots, q_k$ be independent samples from $P_{\mathcal{M}}$, let $\hat{\Phi}_i(r) = \frac{1}{N}\#\{j : d_{\mathcal{M}}(q_i, s_j) \le r\}$ be the empirical LDD of q_i, and let $\varepsilon_N = \sqrt{\ln(2k/\delta)/2N}$. For $t > 0$, let $\hat{m}(t)$ be the smallest $\|\hat{\Phi}_i - \hat{\Phi}_j\|_\infty$ and $m_k(t)$ the smallest $\|\Phi(q_i, \cdot) - \Phi(q_j, \cdot)\|_\infty$, both over pairs with $d_{\mathcal{M}}(q_i, q_j) \ge t$. With probability at least $1 - \delta$, simultaneously for all t,* (i) *every pair with $d_{\mathcal{M}}(q_i, q_j) \ge t$ satisfies $\|\Phi(q_i, \cdot) - \Phi(q_j, \cdot)\|_\infty \ge \hat{m}(t) - 2\varepsilon_N$,* (ii) *$|\hat{m}(t) - m_k(t)| \le 2\varepsilon_N$, and* (iii) *$\hat{m}(t) \le 2\varepsilon_N$ implies $m(t) \le 4\varepsilon_N$.*

Proof. Given q_i, $\hat{\Phi}_i$ is the empirical distribution function of the N i.i.d. distances $d_{\mathcal{M}}(q_i, s_j)$, whose distribution function is $\Phi(q_i, \cdot)$. The DKW inequality with Massart's constant (Theorem A.22; Dvoretzky et al., 1956, Massart, 1990) and a union bound over i give the event E that $\|\hat{\Phi}_i - \Phi(q_i, \cdot)\|_\infty \le \varepsilon_N$ for all i, with probability at least $1 - \delta$. On E, the triangle inequality moves every pairwise distance between LDDs by at most $2\varepsilon_N$, which gives (i). The minima $\hat{m}(t)$ and $m_k(t)$ run over the same pairs, so they also differ by at most $2\varepsilon_N$, and both equal $+\infty$ when no pair is t apart, which gives (ii). If $\hat{m}(t) \le 2\varepsilon_N$, the pair that attains it is at least t apart and has LDDs at most $4\varepsilon_N$ apart, which gives (iii). The event E does not depend on t. □

By (i), the probability that $\hat{m}(t) > 2\varepsilon_N$ while two query points at least t apart share an LDD is at most δ, so rejecting when $\hat{\xi}(t) := \hat{m}(t)/2\varepsilon_N > 1$ is a level-δ test. By (ii), the test detects a margin m among the query points once $4\varepsilon_N < m$, that is, with more than $8\ln(2k/\delta)/m^2$ reference points.

Remark C.6 (Estimation and coverage). Lemma C.5 controls the estimation of each LDD, but not how well the query points cover $\mathcal{M}$. Their margin $m_k(t)$ is an infimum over fewer pairs than $m(t)$, so $m_k(t) \geq m(t)$, and the two meet only as the queries become dense. Whether the test finds a collision therefore depends on how many points share an LDD. Under P_{unif} and P_{vMF}, every point shares its LDD with a continuum of others, random queries land near one of them, and the test does not reject. In the equal-parameter mixture, the collisions above lie on curves and a random pair of queries rarely lies on one, so the test can reject although $m(t) = 0$, which two of the three seeds do at separations from 1.67 on, where few query pairs remain, while the mean in Figure 3c stays below 1. A rejection thus bounds the LDD distances of the queried pairs only.

In low dimension, the queries can cover $\mathcal{M}$, and the estimate then bounds $m(t)$ itself. Under Equation 11, LDDs are Lipschitz in the point, $\|\Phi(x,\cdot) - \Phi(x',\cdot)\|_\infty \leq K\, d_{\mathcal{M}}(x,x')$ (proof of Proposition B.15). If every point of $\mathcal{M}$ lies within γ of a query point, every pair at least t apart lies within γ of a query pair at least $t - 2\gamma$ apart, so on E

$$m(t) \;\geq\; \hat{m}(t - 2\gamma) - 2\varepsilon_N - 2K\gamma.$$

The number of query points such a cover needs grows exponentially with the dimension of $\mathcal{M}$.

The experiments evaluate every LDD on a grid of radii (Appendix E.1). A supremum over the grid is at most the full supremum, and the DKW bound is uniform in r, so everything above holds with the grid norm in place of $\|\cdot\|_\infty$, except that (iii) needs $4\varepsilon_N + Kh$ in place of $4\varepsilon_N$ for radii h apart, since the difference of two LDDs is $2K$-Lipschitz in r.

D The NGMA procedure

NGMA turns the two stages of Section 4 into an estimator. Three things separate it from the idealised procedure analysed there. The latent distances are not given and have to be estimated from the decoders. The coupling comes from a non-convex solver. And the map is a neural network fitted without a single pair, so model selection has to work without pairs too. This appendix gives the details of Section 5 in the same order, and Algorithm 1 summarises the procedure.

The input is computed from the decoders. Following Syrota et al. [2024], each model has an ensemble of K decoders $f_{a,1}, \ldots, f_{a,K}$ on one encoder, and the metric tensor at z is the expected pullback metric

$$G_a(z) = \frac{1}{K} \sum_{k=1}^{K} J_{f_{a,k}}(z)^\top J_{f_{a,k}}(z), \tag{14}$$

which grows where the decoders disagree, that is, away from the data. Geodesic distances require a boundary value problem for every pair, which is out of reach for n^2 pairs. We instead measure each pair by the length of the straight segment between them under Equation 14, discretised into T steps ($T = 32$ for the VAEs and $T = 10$ for the scVI models of Appendix E.6),

$$C^a_{ij} = \sum_{t=1}^{T} \Big(\frac{1}{K} \sum_{k=1}^{K} \big\| f_{a,k}(z_t) - f_{a,k}(z_{t-1}) \big\|^2 \Big)^{1/2}, \qquad z_t = \mathrm{z}^a_i + \tfrac{t}{T}\big(\mathrm{z}^a_j - \mathrm{z}^a_i\big), \tag{15}$$

and likewise for C^b. Both the ensemble metric and the segment length approximate the pullback geometry of Section 2, and the results of Section 4.2 do not cover either approximation. In Appendix E, pullback distances are these segment lengths, except where Appendix E.4 compares them with geodesics.

The coupling is computed by entropic GW [Peyré et al., 2016] on J_β, with C^a and C^b both divided by the mean of their two mean entries, which fixes the overall scale and keeps the relative scale of the two spaces, and with the cost c of Equation 3 computed from the same normalised distances. Four stages lower β through 1, 0.8, 0.5 and 0, and the entropy weight through ε_0, $0.1\varepsilon_0^{2/3}$, $0.01\varepsilon_0^{1/3}$ and 10^{-3}, a geometric interpolation between the LDD-LP's weight $\varepsilon_0 = 0.3 \max c$ and GW's schedule. Each stage starts from the solution of the previous one, and the first from the independent coupling. Each iteration replaces the coupling η by the entropic optimal coupling for the cost $\big(2(1-\beta)L_\eta + \beta c\big)/(2-\beta)$, the derivative of J_β (Appendix B.3.1) divided by $2-\beta$, which for $\beta = 0$ is the update of entropic GW [Peyré et al., 2016]. Each stage runs at most 500 iterations, stopping when the coupling changes by less than 10^{-9} in relative terms, and each iteration solves its transport

problem by log-domain Sinkhorn iterations to a change of 10^{-9}, in double precision. The result is a soft coupling $\hat{\eta} \in \mathbb{R}^{n \times n}$ with uniform marginals, and every later step reads only $\hat{\eta}$, the two samples and their metric tensors. Every stage with $\beta < 1$ has the minimiser $\eta^\star$ under C1 (Proposition B.17), so the anneal changes the path but not the minimiser. At the independent coupling, the linearised GW objective compares points only by their mean distances (Proposition B.18), whereas the first stage compares their whole LDDs. The anneal helps the solver avoid poor stationary points, but it does not guarantee the global one.

Standardisation and targets. Before the maps are fitted, each cloud is centred and divided by its mean norm s_a, and the metric tensors are rescaled to $s_a^2 G_a$ to match the new coordinates, in which the maps act. Standardising removes the offset and scale between the two clouds from what the maps have to learn, and it puts every squared distance in the loss in the same units. The target of z_i^a is the point that $\hat{\eta}$ most likely pairs it with, $\tau_i^a := \mathrm{z}_{j^\star(i)}^b$ with $j^\star(i) = \arg\max_j \hat{\eta}_{ij}$, and symmetrically $\tau_j^b := \mathrm{z}_{i^\star(j)}^a$. This replaces the barycentric projection of the coupling, with which it coincides at $\eta^\star$.

The two maps of Section 5 are

$$\phi(z) = Wz + v + h(z), \qquad \psi(w) = W'w + v' + h'(w), \tag{16}$$

where h and h' are multilayer perceptrons with two hidden layers of width 128 and GELU activations [Hendrycks and Gimpel, 2016], whose last layers are initialised at zero [Zhang et al., 2019]. The affine part starts at the best orthogonal map and translation under the coupling,

$$(Q_0, v_0) = \arg \min_{Q \in O(q),\, v \in \mathbb{R}^q} \sum_{i,j} \hat{\eta}_{ij} \left\| Q\mathrm{z}_i^a + v - \mathrm{z}_j^b \right\|^2, \tag{17}$$

which has a closed form through one singular value decomposition of the $\hat{\eta}$-weighted cross-covariance [Schönemann, 1966]. So $W = Q_0$, $v = v_0$, $W' = Q_0^\top$ and $v' = -Q_0^\top v_0$ at the start, and all parameters are trained, with none shared between the two maps. The coupling is typically right about regions and wrong about individual representations. A rigid map can be fitted from regional information and still extrapolate to every representation, since it has too few parameters to follow the coupling's local mistakes, and everything after this step corrects it.

The objective is Equation 6, with the average over the n_a training points replaced by an average over a batch of 256 of them, and likewise for ℓ_{ba}. Since G_b is known only at the samples, the isometry term reads it at the point $\hat{w}$ of $\mathbf{z}^b$ nearest to $\phi(\mathrm{z}_i^a)$, and no gradient passes through this choice. The isometry term is the pointwise form of $A^* g_b = g_a$ (Proposition 2.3), and it is the only term that reads the geometry. It is also small for maps that send representations to the wrong points with the right local metric, which the fit term penalises by keeping each map close to the coupling's regional structure. The cycle term pushes the two maps to undo each other, as A and A^{-1} do. Measure preservation, the remaining half of an mm-space isomorphism, enters through the uniform marginals of $\hat{\eta}$ and through the selection rule below, which counts one-to-one pairs. Both maps are trained jointly with Adam [Kingma and Ba, 2015] at learning rate 10^{-3} for 3000 steps.

Training and selection without pairs. Before training, we hold out $n_0 = 500$ random points of $\mathbf{z}^a$ and every point of $\mathbf{z}^b$ whose largest entry of $\hat{\eta}$ falls on one of them, so that no held-out point enters the loss. Let $\mathrm{mNN}(\phi; \mathbf{z}^a \to \mathbf{z}^b)$ be the fraction of held-out points z_i^a for which $\phi(\mathrm{z}_i^a)$ and some z_j^b are each other's nearest neighbours, and define $\mathrm{mNN}(\psi; \mathbf{z}^b \to \mathbf{z}^a)$ likewise. At the start and every 250 steps, the pair is scored by

$$s(\phi, \psi) = \tfrac{1}{2}\Big[\mathrm{mNN}\big(\phi;\ \mathbf{z}^a \to \mathbf{z}^b\big) + \mathrm{mNN}\big(\psi;\ \mathbf{z}^b \to \mathbf{z}^a\big)\Big], \tag{18}$$

and the checkpoint with the largest score is kept, the rigid start included. A measure-preserving map between two samples of one distribution induces a near-bijection between them, and mNN counts how much of it a map realises. It reads no pair, and averaging over held-out points keeps it from rewarding a map that memorised its training points. On disjoint samples no map reaches one, so s only compares candidates.

Refinement and ranking. The first fit learns from targets that are wrong for many individual representations, while the pairs that both maps agree on are a smaller and more reliable set. As in unsupervised word translation [Conneau et al., 2018, Artetxe et al., 2018], we refit on those pairs, the mutual nearest neighbours under the combined distance

$$D(i,j) = \omega_a \frac{\|\phi(\mathrm{z}_i^a) - \mathrm{z}_j^b\|}{\mathrm{med}_{i',j'} \|\phi(\mathrm{z}_{i'}^a) - \mathrm{z}_{j'}^b\|} + \omega_b \frac{\|\psi(\mathrm{z}_j^b) - \mathrm{z}_i^a\|}{\mathrm{med}_{i',j'} \|\psi(\mathrm{z}_{j'}^b) - \mathrm{z}_{i'}^a\|}. \tag{19}$$

Algorithm 1 NGMA. The input is two samples of representations and the decoder ensembles of both models. No pair is read.

$C^a, C^b \leftarrow$ segment lengths, Equation 15; $G_a, G_b \leftarrow$ Equation 14 at every sample point
$\hat{\eta} \leftarrow$ entropic GW on J_β, with β lowered from 1 to 0 and the regularisation from ε_0 to 10^{-3}
standardise both clouds; $\tau^a, \tau^b \leftarrow$ the points at the row and column maxima of $\hat{\eta}$
hold out n_0 points of $\mathbf{z}^a$ and the points of $\mathbf{z}^b$ whose largest entry of $\hat{\eta}$ falls on them
$(Q_0, v_0) \leftarrow$ Equation 17; $\phi \leftarrow$ the rigid map, $\psi \leftarrow$ its inverse
train (ϕ, ψ) on Equation 6; keep the checkpoint with the largest s, Equation 18
for up to three rounds **do**
    anchors $\leftarrow$ mutual nearest neighbours under D, Equation 19; replace their targets
    train (ϕ, ψ) on Equation 6 with half of each batch from the anchors
end for
return the round with the largest held-out mutual-nearest-neighbour fraction under D

Each weight is proportional to the squared fraction of points whose image under its map is a mutual nearest neighbour of a point of the other sample, and the two sum to one, so a map that pairs few points contributes little. These mutual nearest neighbours become anchors. For anchors on training points the targets are replaced, $\tau_i^a \leftarrow \mathrm{z}_j^b$ and $\tau_j^b \leftarrow \mathrm{z}_i^a$, half of every batch is drawn from the anchors, and both maps are trained for 1000 steps. We run up to three such rounds, stop early when the number of anchors grows by less than 2%, and keep the round with the largest held-out mutual-nearest-neighbour fraction under D, the state before the first round included. The kept ϕ aligns any new representation of model a, ψ aligns in the other direction, and D ranks candidate pairs between two given samples, which is how test candidates are ranked.

Relation to existing procedures. Procrustes initialisation and refinement on mutual nearest neighbours come from unsupervised word translation [Conneau et al., 2018, Artetxe et al., 2018]. Conneau et al. [2018] also select models without pairs, by the mean similarity of the translations a map induces, where we count mutual pairs on held-out points. Fitting both directions with a cycle term follows Zhu et al. [2017]. GENOT [Klein et al., 2024] also learns a map from GW couplings; other neural GW methods train the map with a Gromov–Monge gap [Sotiropoulou and Alvarez-Melis, 2024, Uscidda et al., 2025] or estimate the entropic GW coupling with neural networks [Wang and Goldfeld, 2026]. NGMA differs from them in what the map has to be, an isometry of the two pullback geometries, which is what the isometry term reads and what Section 2 says the target is.

E Experimental details

E.1 Data on the sphere

Measures. The uniform measure, a vMF measure with concentration $\kappa = 8$, and the mixture P_{mix} of Figure 2: three vMF components with centres $c_1 = (0.556, -0.148, 0.818)$, $c_2 = (0.168, 0.594, 0.787)$, $c_3 = (0.451, 0.804, -0.387)$, concentrations 8, 12, 16 and weights 0.42, 0.33, 0.25. The centres are linearly independent and 49.5°, 75.6° and 100.7° apart, so P_{mix} satisfies C1 by Proposition B.20. The fourth measure is P_{mix} with equal weights and concentration 12. It still satisfies C1, because its centres are linearly independent with distinct pairwise angles, but its LDDs collide along curves that contain the three antipodal pairs $\pm x_k$ with $x_k \propto (c_i + c_j) \times c_k$ (Appendix C), so $m(t) = 0$ at every t. We check both numerically, computing LDDs by one-dimensional quadrature, since the mass of a cap under a vMF component depends only on the angle between the cap and the component's centre. On the curves the LDDs of the equal-parameter mixture agree to within 10^{-10}, while at the antipodal pairs those of P_{mix} differ by 0.03 to 0.16 in sup norm.

Samples and regimes. The alignment is the rotation R by 60° about $(0, 1, 1)/\sqrt{2}$. Each side has $n = 2000$ points, over ten seeds. With independent samples, $\mathbf{z}^a$ and $\mathbf{z}^b$ are drawn independently from P and $R_\# P$. With permuted samples, $\mathrm{z}^b_{\sigma(i)} = R\mathrm{z}^a_i$ for a random permutation σ, and C^b is C^a with rows and columns permuted by σ.

Couplings and landmarks. The LDD-LP and GW are solved on k landmarks per side, chosen greedily by D-optimal design on the LDDs [Pukelsheim, 2006]. LDDs are evaluated on $B = 256$

radii spread evenly over $[0, \pi]$, the same on both sides. Each step of the design adds the point whose LDD most increases $\log\det G_S$, the Gram matrix of the selected LDDs S. Since $\det G_{S\cup\{i\}} = \det G_S \, \|h_i - P_S h_i\|^2$, with h_i the LDD of point i on the grid of radii and P_S the projection onto the span of the selected LDDs, each step adds the LDD farthest from that span. The LDDs span at most B dimensions, so the design orders at most B landmarks. We stop it there, or earlier once the largest increment falls below 10^{-7}, and continue by farthest-point sampling in the L^2 distance between LDDs. Each k takes a prefix of one ordering per side, so in the permutation regime the two sides choose the same points. On learned representations, NGMA solves the coupling on all sample points. The LDD-LP uses the cost of Equation 3 on the geodesic distances from each landmark to all points of its sample, and we solve it exactly by network simplex [Flamary et al., 2021]. In Figure 3b, NGMA's coupling is the annealing of Appendix D on the same landmarks, with the LDD-LP's cost as its linear term, and entropic GW anneals ε over the same four stages without that term. Both cap the inner Sinkhorn at 1000 iterations per step. We run them on the two mixtures only, and NGMA's coupling on independent samples only, since under P_{unif} and P_{vMF} no coupling identifies R. All 960 linear programs reached their optimum, and no stage of the 480 GW solves reached its iteration limit. In one GW coupling (P_{mix}, $k = 128$) the inner Sinkhorn stopped at its iteration limit with column marginals off by up to 10%. Of the 240 solves of NGMA's coupling, 91 reached the limit of 500 iterations in their last one or two stages, mostly at $k = 16$–128; all keep their marginals to within 3×10^{-7}.

Map and error. We fit the rotation in closed form, as the $\hat{R} \in SO(3)$ minimising $\sum_{ij} \eta_{ij} \|\hat{R} z_i^a - z_j^b\|^2$ over the landmark coupling η, and report the angle of $\hat{R}^\top R$.

Margin test. Figure 3c runs the test of Section 4.1 with $N = 512{,}000$ reference points, $k = 2000$ query points, $B = 2000$ radii and $\delta = 0.05$, over three seeds, at 129 separations in $[0, \pi]$. It stops at the largest t at which every measure keeps 1000 query pairs.

E.2 Learned representations: setup

Models. The VAEs on MNIST and FashionMNIST have a four-layer convolutional encoder and residual upsampling decoders, with a 16-dimensional latent space. On CelebA (64×64 images) the latent space has 32 dimensions. Each model has one encoder and an ensemble of eight decoders. The loss is the negative ELBO with a KL weight of 16, a Gaussian likelihood with one learned scale and free bits of 0.25, and the KL weight is warmed up over 100 epochs (30 on CelebA). The prior is a standard Gaussian or a VampPrior [Tomczak and Welling, 2018] with 500 pseudo-inputs (200 on CelebA). We train with Adamax at learning rate 10^{-3}, batch size 128 (256 on CelebA), for 500 epochs (150 on CelebA), and keep the checkpoint with the best validation ELBO.

Pairings and draws. Each dataset has four models per prior. A same-prior setting aligns the six pairs among them, and the cross-prior setting four pairs of a VampPrior and a Gaussian model. On CelebA there are three Gaussian, two VampPrior and two cross-prior pairings. NGMA, Rosetta and mini-vec2vec run five draws per pairing, the other methods at least two.

Samples. Both samples come from the validation split (6000 images on MNIST and FashionMNIST, $15{,}000$ on CelebA). Disjoint samples split it in two halves of $n = 3000$ (7500 on CelebA). The overlap sweep of Figure 6 shares 0, 5, 10, 15, 33, 66 or 100% of the images, with two draws for every MNIST and FashionMNIST pairing.

Latent geometry. The metric of a model is the ensemble pullback metric of Equation 14 with $K = 8$ decoders and exact Jacobians in pixel space [Syrota et al., 2024], and latent distances are the segment lengths of Equation 15 with $T = 32$.

Couplings. NGMA's coupling is $\hat{\eta}$, computed as in Appendix D. The LDD-LP coupling of Figure 6 uses the cost of Equation 3 on the empirical LDDs, divided by its maximum, with Sinkhorn iterations at entropy weight 3×10^{-4}.

Evaluation. Each method maps the test representations of one model onto the other, and we score retrieval among all test points, $10{,}000$ on MNIST and FashionMNIST and $19{,}962$ on CelebA. R@1

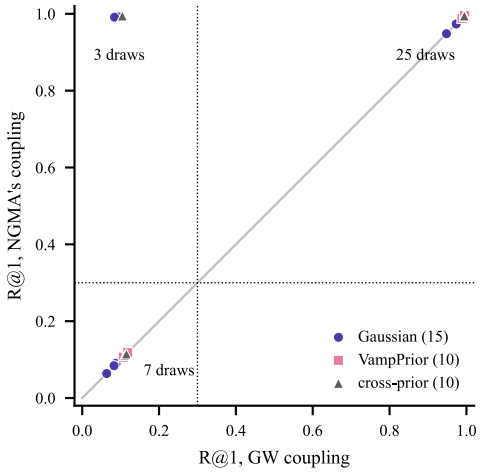


Figure 5: NGMA on CelebA, every draw of every model pairing (35 draws): R@1 from NGMA's coupling against R@1 from entropic GW alone, on the same samples with the same map. Dotted lines mark R@1 0.3. 28 draws reach it with NGMA's coupling and 25 with GW; the coupling recovers 3 draws on which GW fails and loses 0, and no draw is lower by more than 0.004.

is the fraction of queries whose true partner is the nearest candidate, with ties averaged. NGMA ranks the candidates of a query by the combined distance of Equation 19, which weights both maps by the squared fraction of mutual nearest neighbours each one reaches on the samples it was fitted to, so the ranking reads no test pairs. We score every method by this rule. For an orthogonal map (the rigid map, Rosetta, mini-vec2vec and the final rotation of LFM), the inverse preserves distances, and RelReps and RelGeo compare both sides in one shared space. In both cases the $b \to a$ distances are the transpose of the $a \to b$ ones, so the rule gives their $a \to b$ ranking. GENOT-Q's flow has no inverse, and we score it $a \to b$. The NGMA ceiling reports the mean R@1 of its two directions. CSLS@1, R@1 after cross-domain similarity local scaling [Conneau et al., 2018], uses 10 neighbours on each side and the negative Euclidean distance as similarity. For the baselines on CelebA, it is computed on a random subsample of 10,000 test points. Means are over model pairings of each pairing's mean, and the $\pm$ is the s.d. over all draws. A draw with R@1 below 0.3 counts as failed.

Baselines. The two ceilings are fitted on 3000 true pairs (7500 on CelebA) from the validation split. The rigid map is an orthogonal map with a translation, fitted to the pairs by Procrustes [Schönemann, 1966]. The NGMA ceiling is NGMA's map class trained by least squares on the pairs, without NGMA's other loss terms. RelReps [Moschella et al., 2023] represents each point by its cosine similarities to class-centroid anchors, and RelGeo [Yu et al., 2025] by its distances to them, measured as straight-segment lengths under the decoder ensemble. On CelebA each of the 40 attributes gives an anchor at the mean representation of the images that have it. RelReps-Centroid also adds an anchor for the images that lack it. Latent functional maps (LFM) [Fumero et al., 2024] build a 100-nearest-neighbour graph on each sample, fit a functional map between their first 50 Laplacian eigenvectors from heat-kernel signatures, class labels or 50 anchor pairs, and end with a rotation fitted on the matched points. Rosetta [Schnaus et al., 2026] and mini-vec2vec [Dar, 2025] run with their authors' code and defaults on centred, unit-normalised representations. GENOT-Q [Klein et al., 2024] is our implementation: entropic GW with $\varepsilon = 10^{-2}$ on the full samples, with the Euclidean or the pullback cost, followed by a conditional flow that maps each test point to the mean of eight samples. We do not report vec2vec [Jha et al., 2025]. Its run and checkpoint were selected on the test pairs, and it stayed near chance in every setting.

E.3 Learned representations: results

All methods. Tables 2–4 list every method and NGMA variant behind Table 1, with CSLS@1 beside R@1. Figure 5 compares NGMA's coupling with entropic GW alone on every CelebA draw.

Table 2: Instance retrieval on MNIST from disjoint samples ($n = 3{,}000$ per side), every method, scored as in Table 1. Mean $\pm$ s.d. is the mean over model pairings (six for Gaussian and VampPrior, four for cross-prior) of each pairing's mean, with the s.d. over all draws. NGMA, Rosetta and mini-vec2vec use five draws per pairing, the other rows at least two. CSLS@1 is top-1 retrieval under the hubness correction of Conneau et al. [2018]. RelReps uses class-centroid anchors, so RelReps-Centroid is the same computation as RelReps $a \to b$ here and is not repeated. NGMA is ahead of every baseline not given the true pairs in all three columns: 0.967 against 0.941 (Rosetta), 0.989 against 0.930 (Rosetta), 0.955 against 0.571 (Rosetta). With the cross-prior pair, NGMA from GW on the Euclidean cost reaches 0.146 and from the LDD-LP coupling 0.894.

	Gaussian		VampPrior		cross-prior	
Method	R@1	CSLS@1	R@1	CSLS@1	R@1	CSLS@1
Given the true pairs						
NGMA, fitted on the true pairs	0.998±0.001	0.999±0.000	1.000±0.000	1.000±0.000	0.999±0.000	1.000±0.000
Rigid map, fitted on the true pairs	0.955±0.021	0.975±0.014	0.936±0.008	0.954±0.012	0.512±0.021	0.441±0.032
Label-informed (class labels or anchors)						
RelReps	0.525±0.069	0.585±0.069	0.297±0.082	0.338±0.096	0.004±0.001	0.007±0.001
RelGeo	0.512±0.124	0.564±0.131	0.552±0.235	0.603±0.233	0.319±0.152	0.370±0.169
LFM, class labels	0.024±0.011	0.028±0.013	0.100±0.037	0.156±0.054	0.004±0.003	0.007±0.003
LFM, 50 paired anchors	0.040±0.026	0.050±0.035	0.064±0.082	0.100±0.122	0.004±0.002	0.005±0.004
Given no pairs						
LFM, unsupervised	0.000±0.000	0.000±0.000	0.004±0.005	0.006±0.008	0.000±0.000	0.000±0.000
Rosetta	0.941±0.028	0.962±0.021	0.930±0.012	0.952±0.015	0.571±0.021	0.793±0.013
mini-vec2vec	0.849±0.193	0.874±0.186	0.922±0.011	0.945±0.013	0.393±0.180	0.586±0.248
GENOT-Q, Euclidean cost, $a \to b$	0.002±0.003	–	0.076±0.005	–	0.001±0.001	–
GENOT-Q, pullback cost, $a \to b$	0.033±0.004	–	0.088±0.006	–	0.027±0.002	–
NGMA	0.967±0.021	0.986±0.012	0.989±0.012	0.994±0.008	0.955±0.008	0.988±0.005
NGMA, GW coupling	0.963±0.025	0.984±0.015	0.989±0.012	0.994±0.008	0.955±0.006	0.988±0.004
NGMA, LDD-LP coupling	0.741±0.155	0.829±0.150	0.981±0.012	0.991±0.008	0.894±0.019	0.967±0.009
NGMA, GW on the Euclidean cost	0.972±0.023	0.987±0.015	0.984±0.015	0.991±0.012	0.146±0.215	0.179±0.247
NGMA, fully Euclidean	0.949±0.024	0.972±0.016	0.954±0.011	0.972±0.010	0.026±0.034	0.047±0.061

Equivalence and isometry. The two models of a cross-prior pair are far from equivalent (Table 5), yet their pullback distances agree under the true correspondence, while their Euclidean distances do not (Table 6). The table compares the distance between two images in one model with their distance in the other, and counts how often an image has the same nearest neighbour in both models. This is the situation of Proposition 2.3, which needs a shared data manifold, not equivalent models.

Shared images. Figure 6 shows NGMA's R@1 against the share of images the two samples have in common, in every MNIST and FashionMNIST setting. With all images shared, NGMA reaches 0.990–0.999, above the rigid map fitted on the true pairs in every setting. At every share, NGMA has the highest R@1 of the five methods in every setting (Table 7), including the label-informed RelReps and RelGeo. Rosetta and mini-vec2vec centre and normalise the codes before fitting their rotation, so they can exceed that rigid map; Rosetta does in MNIST cross-prior at every share (0.571–0.633 against 0.512).

E.4 Straight segments against geodesics

NGMA measures latent distances by the lengths of straight segments (Equation 15) instead of geodesic distances, which need an optimisation for every pair. Optimising a path for each of the n^2 pairs of a sample is out of reach, so for this comparison we compute geodesics only between landmarks, 500 points per model chosen by the D-optimal design of Appendix E.1, which picks points whose LDDs differ as much as possible. Each geodesic starts from the straight segment between its endpoints, and we optimise its path under Equation 14. Within each model of a Gaussian and a VampPrior pair of FashionMNIST models, the segments are longer than the geodesics by a median of 11.5–15.6% and keep their order, with Spearman correlations of 0.974–0.984.

Comparing the two models of a pair needs distances between the same images in both. Both models of a cross-prior pair encode the same sample of images, but each chooses its landmarks from its own LDDs, so the two choices differ and share only 67 images. On these images and under the true correspondence, the distances of the two models correlate at 0.994 with geodesics and at 0.981 with segments.

Table 3: Instance retrieval on FashionMNIST from disjoint samples ($n = 3{,}000$ per side), every method, scored as in Table 1. Mean $\pm$ s.d. is the mean over model pairings (six for Gaussian and VampPrior, four for cross-prior) of each pairing's mean, with the s.d. over all draws. NGMA, Rosetta and mini-vec2vec use five draws per pairing, the other rows at least two. CSLS@1 is top-1 retrieval under the hubness correction of Conneau et al. [2018]. RelReps uses class-centroid anchors, so RelReps-Centroid is the same computation as RelReps $a \to b$ here and is not repeated. NGMA is ahead of every baseline not given the true pairs in all three columns: 0.719 against 0.710 (Rosetta), 0.979 against 0.656 (Rosetta), 0.871 against 0.098 (RelGeo). With the cross-prior pair, NGMA from GW on the Euclidean cost reaches 0.643 and from the LDD-LP coupling 0.614.

	Gaussian		VampPrior		cross-prior	
Method	R@1	CSLS@1	R@1	CSLS@1	R@1	CSLS@1
Given the true pairs						
NGMA, fitted on the true pairs	0.997±0.001	0.999±0.000	1.000±0.000	1.000±0.000	0.998±0.001	0.999±0.000
Rigid map, fitted on the true pairs	0.782±0.032	0.846±0.031	0.747±0.029	0.847±0.024	0.071±0.004	0.062±0.005
Label-informed (class labels or anchors)						
RelReps	0.151±0.053	0.180±0.059	0.062±0.013	0.073±0.016	0.002±0.001	0.002±0.001
RelGeo	0.159±0.134	0.182±0.151	0.355±0.131	0.394±0.146	0.098±0.033	0.123±0.040
LFM, class labels	0.011±0.004	0.014±0.006	0.020±0.010	0.033±0.016	0.001±0.001	0.002±0.001
LFM, 50 paired anchors	0.060±0.034	0.080±0.046	0.029±0.027	0.051±0.041	0.001±0.001	0.002±0.001
Given no pairs						
LFM, unsupervised	0.001±0.000	0.001±0.001	0.000±0.000	0.001±0.001	0.000±0.000	0.000±0.000
Rosetta	0.710±0.040	0.761±0.037	0.656±0.045	0.762±0.044	0.065±0.019	0.178±0.047
mini-vec2vec	0.504±0.274	0.547±0.292	0.624±0.030	0.738±0.028	0.018±0.010	0.044±0.026
GENOT-Q, Euclidean cost, $a \to b$	0.007±0.005	–	0.021±0.012	–	0.002±0.002	–
GENOT-Q, pullback cost, $a \to b$	0.009±0.001	–	0.030±0.002	–	0.006±0.001	–
NGMA	0.719±0.112	0.812±0.107	0.979±0.027	0.989±0.017	0.871±0.049	0.965±0.017
NGMA, GW coupling	0.733±0.107	0.824±0.102	0.982±0.023	0.990±0.015	0.872±0.063	0.964±0.024
NGMA, LDD-LP coupling	0.268±0.097	0.340±0.117	0.953±0.065	0.971±0.054	0.614±0.076	0.797±0.072
NGMA, GW on the Euclidean cost	0.561±0.317	0.622±0.330	0.917±0.198	0.929±0.189	0.643±0.329	0.738±0.377
NGMA, fully Euclidean	0.536±0.284	0.593±0.297	0.815±0.217	0.862±0.212	0.051±0.029	0.078±0.045

E.5 The CelebA failures are mirror alignments

Seven CelebA draws fail with NGMA's coupling and ten with GW alone (Table 1, Figure 5). A left-right flip permutes pixels, so it preserves distances in the observation space. Since it maps faces to faces, its restriction to the data manifold preserves $d_{\mathcal{M}}$, and since a face and its mirror image are about equally likely, it nearly preserves $P_{\mathcal{M}}$. The flip is then nearly a symmetry of the data mm-space, $f_b^{-1} \circ \mathrm{flip} \circ f_a$ is nearly an isomorphism of the latent mm-spaces, and C1 nearly fails.

We test this directly. Every test image x is flipped into x', and both models encode the flipped images. If a fitted map follows the mirror alignment, it sends the representation of x in one model close to the representation of x' in the other. True R@1 scores each map against the representations of the unflipped images, and mirror R@1 against those of the flipped images with x' as the partner, in each direction separately. The self-mirror rate of a model is the fraction of images whose own mirror is the nearest flipped representation, and it measures how far the model keeps a face and its mirror apart. We score every reported CelebA draw with both couplings, and as controls the Gaussian pairing ab of MNIST, whose digits are not mirror-symmetric, and of FashionMNIST, whose garments often are.

On every failed draw, with either coupling and in either direction, mirror R@1 is 0.870–0.993 and true R@1 at most 0.113, and on every successful draw the reverse holds (Table 8). The models keep a face and its mirror apart, since a face's own mirror is the nearest flipped representation for only 10–12% of faces. A high mirror R@1 therefore means that the map sends faces to their mirrors, not that the models confuse the two. On MNIST, mirror R@1 stays below 0.01, and on FashionMNIST it lies in between. The mirror alignment also explains why the objectives cannot tell failed draws from successful ones. The flip preserves distances and nearly preserves LDDs, so GW and the LDD-LP score the mirror alignment almost as well as A, and on the GW couplings the GW objective is 0.0103–0.0108 on failed draws against 0.0104–0.0106 on successful ones.

E.6 scVI models of lung cells

Data. We use the $62{,}954$ cells of the Tabula Sapiens lung measured with 10x $3'$ chemistry [The Tabula Sapiens Consortium, 2022], from four donors with 34 annotated cell types, on the 3000 most

Table 4: Instance retrieval on CelebA from disjoint samples ($n = 7{,}500$ per side), every method, scored as in Table 1. Mean $\pm$ s.d. is the mean over model pairings (three for Gaussian, two for VampPrior and two for cross-prior) of each pairing's mean, with the s.d. over all draws. NGMA, Rosetta and mini-vec2vec use five draws per pairing, the other rows at least two. CSLS@1 is top-1 retrieval under the hubness correction of Conneau et al. [2018]. NGMA is ahead of every baseline not given the true pairs in all three columns: 0.804 against 0.567 (Rosetta), 0.818 against 0.616 (mini-vec2vec), 0.818 against 0.757 (Rosetta). With the cross-prior pair, NGMA from GW on the Euclidean cost reaches 0.006 and from the LDD-LP coupling 0.643. Where a draw of NGMA has R@1 below 0.3, the parentheses count the draws that reach it. From NGMA's coupling, 28 of the 35 draws reach it; from GW alone, 25.

	Gaussian		VampPrior		cross-prior	
Method	R@1	CSLS@1	R@1	CSLS@1	R@1	CSLS@1
Given the true pairs						
NGMA, fitted on the true pairs	0.995±0.000	0.995±0.000	0.996±0.000	0.997±0.000	0.996±0.000	0.996±0.000
Rigid map, fitted on the true pairs	0.993±0.000	0.997±0.000	0.995±0.000	0.998±0.000	0.990±0.000	0.996±0.000
Label-informed (class labels or anchors)						
RelReps-Centroid	0.270±0.028	0.352±0.031	0.249±0.065	0.305±0.077	0.000±0.000	0.000±0.000
RelReps	0.445±0.033	0.575±0.030	0.282±0.070	0.352±0.084	0.000±0.000	0.000±0.000
RelGeo	0.340±0.118	0.424±0.150	0.387±0.139	0.479±0.166	0.220±0.120	0.278±0.151
LFM, class labels	0.060±0.015	0.131±0.033	0.024±0.008	0.068±0.019	0.003±0.002	0.012±0.005
LFM, 50 paired anchors	0.042±0.034	0.100±0.075	0.108±0.092	0.247±0.188	0.006±0.004	0.023±0.013
Given no pairs						
LFM, unsupervised	0.000±0.000	0.001±0.001	0.000±0.000	0.001±0.000	0.000±0.000	0.000±0.000
Rosetta	0.567±0.453	0.586±0.444	0.365±0.434	0.388±0.421	0.757±0.372	0.795±0.367
mini-vec2vec	0.251±0.353	0.278±0.360	0.616±0.415	0.648±0.400	0.439±0.474	0.475±0.482
GENOT-Q, Euclidean cost, $a \to b$	0.000±0.000	–	0.003±0.002	–	0.000±0.000	–
GENOT-Q, pullback cost, $a \to b$	0.001±0.000	–	0.004±0.002	–	0.001±0.001	–
NGMA	0.804±0.375 (12/15)	0.814±0.367	0.818±0.372 (8/10)	0.823±0.363	0.818±0.372 (8/10)	0.823±0.363
NGMA, GW coupling	0.684±0.440 (10/15)	0.697±0.430	0.818±0.372 (8/10)	0.823±0.362	0.729±0.427 (7/10)	0.736±0.417
NGMA, LDD-LP coupling	0.467±0.434	0.515±0.424	0.733±0.328	0.759±0.311	0.643±0.378	0.670±0.362
NGMA, GW on the Euclidean cost	0.424±0.446	0.458±0.442	0.464±0.457	0.478±0.446	0.006±0.018	0.010±0.028
NGMA, fully Euclidean	0.387±0.430	0.423±0.420	0.451±0.464	0.468±0.452	0.000±0.001	0.001±0.002

Table 5: Likelihood and generated distributions of the two models of every pair: the difference of their importance-weighted (IWAE) bounds on 2000 test images ($K = 1000$ importance samples), and the squared maximum mean discrepancy (MMD^2) between their generated images, and of each against the data, with the permutation p-value of the first. Ranges over pairings; in the cross-prior settings a is the VampPrior model. The cross-prior pairs differ by 10.2–25.2 nats, the same-prior pairs by at most 4.3. The two generators of a cross-prior pair are further apart than the VampPrior model is from the data on 7 of 8 pairs, and than the Gaussian model is from the data on 0. The MMD test also tells 18 of 24 same-prior pairs apart ($p < 0.05$; 0.005 is the smallest p that 200 permutations can return), so the IWAE gap, not the MMD, is what sets the cross-prior pairs apart. They are isometric (Table 6) without being equivalent, as Proposition 2.3 allows.

		IWAE (nats)		Generator MMD^2 ($\times 10^3$)			
Setting	pairs	$a - b$	s.e.	a vs. b	a vs. data	b vs. data	p
MNIST, Gaussian	6	-4.3–$+0.5$	0.6–0.8	0.4–1.3	8.0–10.5	8.1–8.7	0.005–0.010
MNIST, VampPrior	6	-0.6–$+3.9$	0.6–0.8	0.8–1.6	2.5–2.7	2.3–2.9	0.005
MNIST, cross-prior	4	$+22.3$–$+25.2$	0.7–0.9	2.6–5.2	2.5–2.7	7.4–8.7	0.005
FashionMNIST, Gaussian	6	-1.0–$+1.6$	0.3–0.4	0.5–1.2	13.7–16.5	12.7–13.7	0.005–0.055
FashionMNIST, VampPrior	6	-1.1–$+0.2$	0.3–0.4	0.0–0.7	3.1–3.3	2.6–3.0	0.030–0.428
FashionMNIST, cross-prior	4	$+10.2$–$+11.1$	0.5	6.2–7.3	3.1–3.3	12.7–13.2	0.005

variable genes, as raw counts. The cells are split at random into $10{,}000$ held-out evaluation cells, a pool of 9000 cells from which the samples are drawn, and two disjoint training halves, each with nested subsets of 4900, 9800 and $19{,}600$ cells. Every donor appears in every part.

Models and geometry. Each model is scVI [Lopez et al., 2018] with the recipe of the published Tabula Sapiens lung model: a 20-dimensional latent space, a negative-binomial likelihood with one dispersion θ_g per gene, and 400 epochs. We fit one model per training subset and seed. Models on different halves see different cells. We keep each encoder and add seven decoders, trained on the model's own posterior representations with its likelihood. The data metric is the Fisher metric of the negative-binomial mean, which is Euclidean in the coordinates $y_g = 2\sqrt{\theta_g}\,\operatorname{asinh}\left(\sqrt{l\rho_g/\theta_g}\right)$ at the

Table 6: Pairwise distances in one model against the same pairs in the other, under the true correspondence, on the samples with 100% overlap (one draw, 1500 representations). Pullback distances are segment lengths under the decoder ensemble (Equation 15). *corr.* is the Pearson correlation of the two distance matrices, *scale* the factor by which one exceeds the other, and *same neighbour* the share of images whose nearest neighbour is the same image in both models. Ranges are over pairings. Pullback distances agree across all 32 pairs (corr. 0.973–0.996, scale 0.98–1.01). Euclidean distances do not on the cross-prior pairs, with corr. 0.71, 0.61 and scale 2.92, 5.01 (MNIST, FashionMNIST). A random pairing gives $|\text{corr.}| \leq 0.01$.

		Pullback distances			Euclidean distances		
Setting	pairs	corr.	scale	same neighbour	corr.	scale	same neighbour
MNIST, Gaussian	6	0.987–0.991	0.98–0.99	0.84–0.87	0.903–0.941	0.99–1.00	0.58–0.63
MNIST, VampPrior	6	0.992–0.994	0.99–1.00	0.86–0.90	0.961–0.973	0.98–1.02	0.74–0.75
MNIST, cross-prior	4	0.976–0.983	1.00–1.01	0.84–0.86	0.696–0.714	2.89–2.95	0.43–0.44
FashionMNIST, Gaussian	6	0.979–0.989	1.00	0.81–0.84	0.905–0.928	0.99–1.00	0.57–0.60
FashionMNIST, VampPrior	6	0.994–0.996	0.98–1.01	0.85–0.86	0.972–0.978	1.00–1.02	0.72–0.77
FashionMNIST, cross-prior	4	0.973–0.976	0.99–1.00	0.79–0.81	0.598–0.623	4.98–5.04	0.31–0.33

median library size l, and latent distances are the segment lengths of Equation 15 with $T = 10$ under the eight decoders.

Pairs, samples and evaluation. For each training size we align two pairs of models trained on different cells and two trained on the same cells, on disjoint samples of 3000 cells per side and three draws per pair, with the settings of Appendix D. Both models encode the same $10{,}000$ evaluation cells, which gives the true pairs. We report R@1 among these cells (chance 10^{-4}) and the cell-type accuracy of the nearest cell, both scored by the rule of Appendix E.2.

Results. NGMA is above every baseline on all 36 samples, by 0.109–0.298 in R@1, and above the rigid map fitted on the true pairs on all of them (Table 9). With Euclidean geometry, R@1 stays below 0.01 on 10 of the 18 draws between models trained on different cells, and with the pullback metric on none.

Single cells are only partly determined by these models. The map fitted on the true pairs reaches 0.36–0.69, and NGMA 59–72% of it.

Limits. Both models read the same 3000 genes, so any cell can be encoded by both and pairs are not scarce. The experiment tests whether the two models alone identify the alignment, not a use in biology. The models have no batch covariate, so their representations carry donor structure, which both sides share under the random split.

E.7 Compute

Every GPU job ran on one graphics compute die (GCD), with 64 GB of device memory and 40 GB of host memory, and Table 10 gives the time of each stage for the image experiments. The pullback distances take nearly all of it. They decode the segment between every pair of points under eight decoders, so their cost grows as n^2; at CelebA's $n = 15000$ they take about 810 GPU-hours per model. The sphere experiments run on a CPU.

Table 7: R@1 against the share of images the two samples have in common, every method on every MNIST and FashionMNIST pairing: the mean over a setting's pairings of the mean over draws. At 0% the samples are the disjoint samples of Table 1, with five draws per pairing (two for RelReps and RelGeo); every other share has two draws per pairing. RelReps and RelGeo are label-informed. Rosetta and mini-vec2vec use no pairs and no labels, like NGMA. The rigid map fitted on the true pairs does not depend on the share, so each setting's header gives its R@1. Rosetta and mini-vec2vec centre and normalise the codes before fitting their rotation, so they can exceed it. The best R@1 in each column of a setting is in bold. NGMA has the highest R@1 in every column of every setting. On single pairings, a baseline matches or exceeds NGMA at 4 of 896 points (method, pairing, share): Rosetta on FashionMNIST, Gaussian ac at 0% (0.652 against 0.537); Rosetta on MNIST, Gaussian ac at 5% (0.921 against 0.917); Rosetta on MNIST, Gaussian ac at 10% (0.926 against 0.919); Rosetta on MNIST, Gaussian ad at 15% (0.934 against 0.932).

	Share of images in common (%)						
Method	0	5	10	15	33	66	100
MNIST, Gaussian; rigid map on the true pairs 0.955							
NGMA	**0.967**	**0.963**	**0.969**	**0.982**	**0.989**	**0.999**	**0.999**
RelReps	0.525	0.531	0.531	0.539	0.549	0.573	0.600
RelGeo	0.512	0.491	0.485	0.482	0.451	0.496	0.476
Rosetta	0.941	0.945	0.950	0.952	0.956	0.957	0.958
mini-vec2vec	0.849	0.805	0.861	0.792	0.862	0.887	0.945
MNIST, VampPrior; rigid map on the true pairs 0.936							
NGMA	**0.989**	**0.992**	**0.988**	**0.988**	**0.997**	**0.998**	**0.999**
RelReps	0.297	0.299	0.300	0.299	0.301	0.300	0.305
RelGeo	0.552	0.665	0.673	0.661	0.672	0.722	0.759
Rosetta	0.930	0.927	0.929	0.929	0.931	0.930	0.931
mini-vec2vec	0.922	0.921	0.922	0.922	0.925	0.926	0.927
MNIST, cross-prior; rigid map on the true pairs 0.512							
NGMA	**0.955**	**0.967**	**0.969**	**0.975**	**0.985**	**0.988**	**0.998**
RelReps	0.004	0.003	0.003	0.003	0.002	0.003	0.004
RelGeo	0.319	0.351	0.353	0.342	0.349	0.347	0.385
Rosetta	0.571	0.581	0.590	0.597	0.615	0.626	0.633
mini-vec2vec	0.393	0.463	0.460	0.407	0.520	0.506	0.534
FashionMNIST, Gaussian; rigid map on the true pairs 0.782							
NGMA	**0.719**	**0.854**	**0.883**	**0.928**	**0.986**	**0.996**	**0.998**
RelReps	0.151	0.160	0.161	0.161	0.164	0.172	0.186
RelGeo	0.159	0.148	0.154	0.153	0.159	0.165	0.193
Rosetta	0.710	0.724	0.728	0.732	0.745	0.756	0.764
mini-vec2vec	0.504	0.663	0.582	0.629	0.586	0.652	0.699
FashionMNIST, VampPrior; rigid map on the true pairs 0.747							
NGMA	**0.979**	**0.995**	**0.996**	**0.996**	**0.997**	**0.999**	**0.999**
RelReps	0.062	0.062	0.061	0.063	0.065	0.068	0.066
RelGeo	0.355	0.354	0.363	0.369	0.385	0.411	0.435
Rosetta	0.656	0.685	0.689	0.691	0.700	0.705	0.712
mini-vec2vec	0.624	0.639	0.644	0.650	0.662	0.669	0.680
FashionMNIST, cross-prior; rigid map on the true pairs 0.071							
NGMA	**0.871**	**0.919**	**0.894**	**0.941**	**0.967**	**0.984**	**0.990**
RelReps	0.002	0.002	0.002	0.002	0.002	0.002	0.002
RelGeo	0.098	0.107	0.101	0.101	0.097	0.104	0.082
Rosetta	0.065	0.073	0.074	0.074	0.079	0.081	0.085
mini-vec2vec	0.018	0.016	0.013	0.016	0.019	0.019	0.025

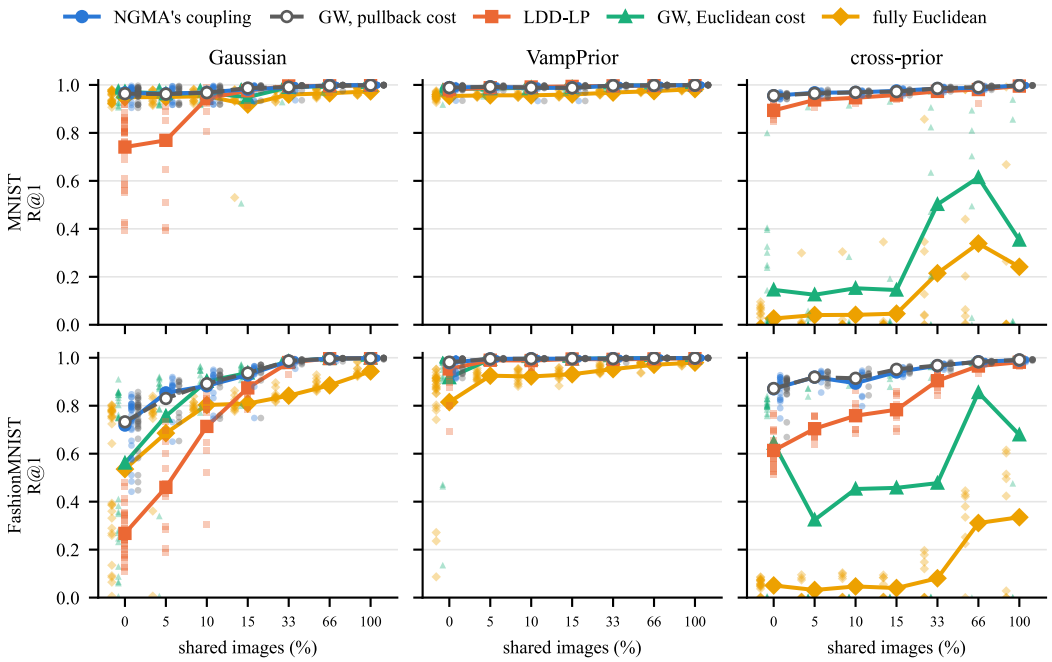


Figure 6: NGMA's R@1 against the share of images the two samples have in common, fitted from five couplings, every MNIST and FashionMNIST setting. Lines join the means over pairings and draws, and dots are single draws. At 0% the samples are the disjoint samples of Table 1, with five draws per pairing; every other share has two. The three GW couplings are plain entropic GW, without NGMA's LDD-LP stages. With all images shared, NGMA's coupling reaches 0.990–0.999 and GW on the pullback cost 0.991–0.999 in every setting, and GW on the Euclidean cost reaches 0.353 on the MNIST cross-prior pair and 0.679 on the FashionMNIST cross-prior pair. The LDD-LP needs shared images where LDDs are poorly separated, as on FashionMNIST with Gaussian priors (0.268 at 0%, 0.874 at 15% and 0.980 at 33%).

Table 8: The mirror test. Every test image is flipped left to right and encoded by both models. True R@1 scores the image of a representation under a map against the representations of the unflipped images, and mirror R@1 against those of the flipped images, with the image's own mirror as the partner. The self-mirror rate is the fraction of images whose own mirror is the nearest flipped representation in the same model. Ranges are over draws. The 70 CelebA alignments are the seven pairings with draws 1–5, each from NGMA's coupling and from GW alone, and a draw fails when its R@1 is below 0.3. The controls are MNIST and FashionMNIST, Gaussian pairing ab, draws 1–5.

		$a \to b$		$b \to a$		
Draws	n	true R@1	mirror R@1	true R@1	mirror R@1	self-mirror rate
CelebA, failed draws						
NGMA's coupling	7	0.054–0.113	0.905–0.992	0.055–0.111	0.910–0.992	0.101–0.116
GW alone	10	0.052–0.113	0.870–0.993	0.053–0.111	0.893–0.993	0.101–0.116
CelebA, successful draws						
NGMA's coupling	28	0.948–0.995	0.057–0.106	0.947–0.994	0.057–0.103	0.101–0.116
GW alone	25	0.948–0.995	0.057–0.107	0.947–0.994	0.057–0.103	0.101–0.116
Controls, both couplings						
MNIST, Gaussian	10	0.873–0.916	0.004–0.005	0.891–0.922	0.005–0.006	0.007
FashionMNIST, Gaussian	10	0.595–0.745	0.155–0.210	0.615–0.760	0.176–0.213	0.276–0.305

Table 9: scVI models of Tabula Sapiens lung cells (Appendix E.6): NGMA against the unpaired baselines, on the disjoint samples ($n = 3000$ cells per side) of two models trained on different cells or on the same cells with different seeds, by training size. Retrieval of the true partner among the $10{,}000$ evaluation cells (R@1, chance 10^{-4}) and cell-type accuracy of the nearest cell (chance 0.124), scored as in Table 1. GENOT-Q runs on its own GW coupling with the Euclidean cost or with the pullback cost NGMA uses. NGMA, one decoder, reads the pullback metric of the scVI decoder alone instead of the ensemble of eight. The last two rows of each block are not methods. They are NGMA's map class and the rigid map, both fitted on the true pairs and scored $a \to b$. Mean $\pm$ s.d. over two pairings and three draws; the best method in each column is in bold. †RelGeo builds its anchors from the cell-type labels of both models.

	different cells			same cells		
Training size	4,900	9,800	19,600	4,900	9,800	19,600
R@1						
NGMA	**0.221**±0.035	**0.369**±0.026	**0.378**±0.015	**0.274**±0.023	**0.401**±0.031	**0.480**±0.016
NGMA, Euclidean	0.119±0.064	0.002±0.002	0.047±0.066	0.062±0.058	0.181±0.118	0.146±0.146
NGMA, one decoder	0.195±0.031	0.348±0.020	0.294±0.075	0.257±0.013	0.372±0.050	0.389±0.073
unpaired baselines						
Rosetta	0.077±0.033	0.141±0.020	0.128±0.011	0.088±0.024	0.154±0.025	0.210±0.010
mini-vec2vec	0.028±0.021	0.062±0.065	0.060±0.039	0.027±0.035	0.071±0.063	0.124±0.027
GENOT-Q, Eucl. cost	0.028±0.019	0.001±0.001	0.002±0.003	0.020±0.016	0.007±0.010	0.015±0.015
GENOT-Q, pullback cost	0.024±0.007	0.030±0.001	0.027±0.001	0.028±0.002	0.033±0.005	0.032±0.002
label-informed						
RelGeo†	0.027±0.005	0.045±0.012	0.047±0.013	0.036±0.005	0.059±0.009	0.051±0.024
fitted on the true pairs						
map	0.363±0.036	0.510±0.005	0.624±0.009	0.466±0.019	0.621±0.011	0.692±0.007
rigid map	0.099±0.029	0.185±0.006	0.181±0.003	0.110±0.020	0.179±0.014	0.254±0.007
cell-type accuracy						
NGMA	**0.891**±0.023	**0.924**±0.008	**0.931**±0.006	0.893±0.011	0.911±0.016	**0.938**±0.014
NGMA, Euclidean	0.701±0.305	0.062±0.052	0.449±0.278	0.567±0.239	0.667±0.209	0.443±0.423
NGMA, one decoder	0.883±0.025	0.923±0.006	0.897±0.046	**0.897**±0.008	**0.914**±0.019	0.915±0.034
unpaired baselines						
Rosetta	0.773±0.028	0.775±0.044	0.789±0.015	0.766±0.048	0.774±0.008	0.814±0.019
mini-vec2vec	0.563±0.169	0.577±0.205	0.636±0.155	0.487±0.245	0.497±0.288	0.727±0.079
GENOT-Q, Eucl. cost	0.633±0.193	0.144±0.100	0.346±0.282	0.551±0.240	0.337±0.271	0.397±0.350
GENOT-Q, pullback cost	0.777±0.014	0.774±0.018	0.768±0.003	0.773±0.003	0.771±0.015	0.768±0.007
label-informed						
RelGeo†	0.825±0.013	0.848±0.027	0.852±0.011	0.845±0.011	0.858±0.008	0.854±0.007
fitted on the true pairs						
map	0.935±0.006	0.950±0.001	0.960±0.003	0.943±0.002	0.960±0.002	0.968±0.001
rigid map	0.793±0.025	0.800±0.035	0.820±0.010	0.802±0.017	0.792±0.006	0.825±0.019

Table 10: Compute per stage, in hours on one GPU: one GCD of an AMD Instinct MI250X, which has two and which the software sees as two GPUs. Distances are pullback distances under the 8-decoder ensemble between all n images a model encodes ($n = 6000$ for MNIST and FashionMNIST, $n = 15000$ for CelebA, of which each draw uses 7500); jobs split by rows compute each pair twice, the upper end of the MNIST/FashionMNIST range. The whole project, development included, used about 50 000 GPU-hours.

Stage	MNIST, FashionMNIST	CelebA
VAE training, per model	3	16–25
Pullback distances, per model	9–18	810
Coupling, per draw	0.02–0.2	0.05–2.3
Map fit and evaluation, per draw	0.1–0.25	0.35–0.75
One pair of models, one draw	25–43	1 700