PLOS Computational Biology, 2026, 1–12
doi: https://doi.org/10.1101/2024.10.23.619602

Signature Distance: Generalizing Energy Statistics

Nicolo' Lazzaro,[1,2,*] Raffaele Marchesi,[1,3] Gianluca Leonardi,[1,4] Jacopo Tessadori,[1] Marco Chierici,[1] Monica Moroni,[1] Gabriele Sales,[5] Toma Tebaldi[4,†] and Giuseppe Jurman[1,6,†]

[1]Fondazione Bruno Kessler, Data Science for Health, Trento, Italy, [2]University of Trento, Department of Information Engineering and Computer Science, Trento, Italy, [3]University of Pavia, Department of Mathematics, Pavia, Italy, [4]University of Trento, Department of Cellular, Computational and Integrative Biology, Trento, Italy, [5]University of Padova, Padova, Italy and [6]Humanitas University, Department of Biomedical Sciences, Milan, Italy

*Corresponding author. nlazzaro@fbk.eu

Abstract

Comparing empirical distributions is central to generative model evaluation, hypothesis testing and data augmentation in high-dimensional biological data. Established methods such as energy distance summarize each point's relationship to the opposing distribution through a single expected distance, providing sensitivity to location shifts but not to local density or topological structure. We introduce Signature Distance (SD), a metric that compares empirical distributions through the mean absolute difference of their sorted pointwise distance profiles. SD is a structural generalization of energy distance and matches its $\mathcal{O}(n^2)$ computational complexity. On TCGA pan-cancer transcriptomic data, we show that (1) SD detects density changes that energy distance is insensitive to; (2) the per-point SD loss landscape reveals the geometric mechanisms behind known limitations of energy distance as a generative objective; (3) linearly interpolated biological samples that are not detected by energy distance are correctly penalized by SD; (4) SD provides a direct differentiable potential energy for model-free Langevin data expansion, with a bootstrap resampling protocol that stabilises the stopping epoch; and (5) SD is directly usable as a differentiable generative training loss. Code to reproduce all experiments is available at `https://github.com/lazzaronico/signature-distance`.

Key words: statistical distance, two-sample testing, generative model validation, optimal transport, energy distance, omics

Introduction

Comparing empirical distributions is a recurring problem in computational biology: assessing whether synthetic data resembles real data (Luecken and Theis, 2019), testing for differences between treatment and control groups, and guiding iterative data generation. In high-dimensional settings, where biological measurements routinely span hundreds to thousands of features, distances between points concentrate (Beyer et al., 1999), and standard comparison methods lose discriminative power.

A widely adopted family of methods relies on expected pairwise distances. Energy distance (ED; Székely and Rizzo (2013)) compares distributions through the difference between the expected cross-distance and the expected within-set distances, yielding a squared distance that embeds in a Hilbert space and supports permutation-based hypothesis testing. Related methods share the same scalar-reduction structure: maximum mean discrepancy (MMD; Gretton et al. (2012)) generalises ED to arbitrary reproducing kernels (with a Euclidean kernel it reduces to ED; Székely and Rizzo 2013), while the Fréchet Inception Distance (FID; Heusel et al. (2017)) compares Gaussian-fitted moments in a learned feature space. We use ED as the representative baseline throughout.

Because this scalar reduction captures only the mean pairwise distance, it is well-suited to detecting global location shifts but insensitive to local geometric structure. Two distributions can have identical expected cross-distances while differing in density, shape, or internal manifold organization. As we show in Section 3.1, controlled density perturbations invisible to ED are detected by SD.

At the other end of the computational spectrum, the d-dimensional Wasserstein distance (W_d; Villani 2009; Peyré and Cuturi 2019) provides a principled framework for geometric distributional comparison, but its $\mathcal{O}(n^3 \log n)$ complexity is

[2] Joint senior authors.

© The Author 2026. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

impractical at the sample sizes common in omics data. Topological data analysis (Carlsson, 2009; Edelsbrunner and Harer, 2010) captures multi-scale geometric structure via filtrations but produces summaries that are expensive to compare natively.

We introduce Signature Distance, which bridges this gap. A point's sorted distances to all members of a population form a one-dimensional fingerprint of its local neighborhood: the nearest neighbors determine the front of the sorted array, while the distant points fill its tail. Because this fingerprint encodes the full radial density profile, it captures structural features such as density gradients, cluster boundaries and manifold curvature that are lost when neighborhood relationships are collapsed to a scalar mean. SD compares these fingerprints across two populations by computing the W_1 distance between each point's intra- and cross-distribution sorted profiles, then symmetrizes over all reference points weighted by the barycentric mixture (Figure 1; see Methods, Section 2.1). This decomposition into one-dimensional W_1 problems reduces computational complexity to $\mathcal{O}(n^2)$, matching ED, while retaining sensitivity to local density structure.

The present work makes five contributions. First, we formally define Signature Distance and establish its metric properties, analytically deriving its structural relationship to ED (Section 2.1). Second, we show that SD detects distributional differences that ED is insensitive to, using controlled perturbations (Section 3.1). Third, we show that per-point SD penalties reveal the geometric mechanisms behind known limitations of ED as a generative objective (Section 3.2). Fourth, we show that linearly interpolated biological samples undetected by ED are correctly penalized by SD, and that SD serves as a differentiable potential energy for model-free Langevin data expansion (Sections 3.3–3.4). Fifth, we show that SD is directly usable as a differentiable generative training loss, and introduce a glocal training protocol for multi-population settings that enables fair comparison of distributional losses on tissue-conditioned generation of TCGA landmark gene expression profiles (Section 3.5).

Methods

Signature Distance

Given two empirical point clouds $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_m\}$ in $\mathbb{R}^d$, Signature Distance is computed in four steps.

First, the pairwise distance matrices are computed:

$$D_{XX}[i,j] = \|x_i - x_j\|, \quad D_{XY}[i,j] = \|x_i - y_j\|, \quad D_{YY}[i,j] = \|y_i - y_j\| \tag{1}$$

Second, for each point $x_i \in X$, we define its *intra-signature* as the sorted i-th row of D_{XX} and its *cross-signature* as the sorted i-th row of D_{XY}:

$$S_{XX}(x_i) = \mathrm{sort}(D_{XX}[i,:]), \quad S_{XY}(x_i) = \mathrm{sort}(D_{XY}[i,:]) \tag{2}$$

When $n \neq m$, the longer signature vector is resampled to match the shorter by linearly interpolating at evenly spaced quantile positions, equivalent to aligning their empirical cumulative distribution functions (eCDFs). This quantile-alignment step is a close approximation to exact W_1 computation; for settings where strict metric guarantees at unequal sample sizes are required, an exact implementation can replace the summation in Equation 3 with the integral $\int_0^\infty |F_1(r) - F_2(r)|\, dr$ over the eCDFs.

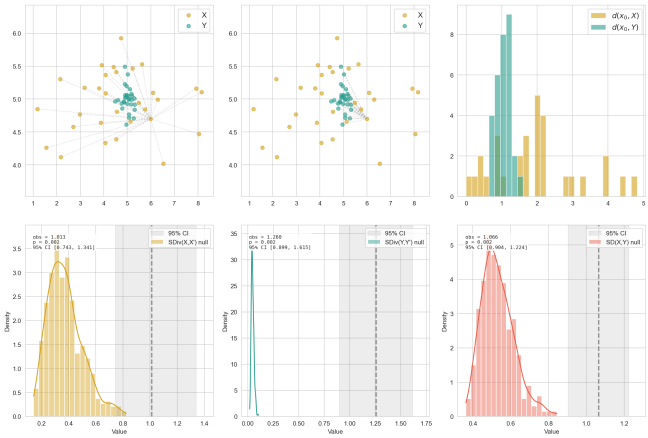

Fig. 1: The geometric signature and its use as a two-sample test. Two populations X (gold, dispersed) and Y (teal, compact cluster) are drawn from different distributions in $\mathbb{R}^2$. *Top left:* the signature of a reference point $x_0 \in X$ consists of its sorted distances to all points in X (intra-distribution distances, dashed lines). *Top centre:* the same reference point's distances to all points in Y (cross-distribution distances). *Top right:* histogram comparison: if X and Y were drawn from the same distribution, these two distance profiles would be statistically similar; their divergence (measured by sorted-rank L1 distance) quantifies how much the neighbourhood geometry around x_0 differs between populations. *Bottom row:* permutation-test null distributions (shaded) and observed statistics (dashed line) for both directional divergences $\mathrm{SDiv}(X \,\|\, Y)$ and $\mathrm{SDiv}(Y \,\|\, X)$, and the symmetric $\mathrm{SD}(X,Y)$. All three exceed the null ($p = 0.002$), confirming that SD detects the density difference between the dispersed and concentrated populations.

Third, the signatures are compared using the mean absolute difference of their sorted quantiles. Because the signatures are one-dimensional sorted arrays, this operation is exactly equivalent to the 1-Wasserstein (W_1) or Earth Mover's distance. We formally define this pointwise *divergence* between two signatures as:

$$\delta(S_1, S_2) = \frac{1}{K} \sum_{k=1}^{K} |S_1[k] - S_2[k]| \equiv W_1(S_1, S_2) \tag{3}$$

Fourth, we reserve the term *distance* exclusively for the symmetric aggregate metric. We define the **Squared Signature Distance (SD^2)** as the symmetrized average pointwise divergence over all points in both distributions:

$$\mathrm{SD}^2(X,Y) = \frac{1}{2n} \sum_{i=1}^{n} \delta\big(S_{XX}(x_i),\, S_{XY}(x_i)\big) + \frac{1}{2m} \sum_{j=1}^{m} \delta\big(S_{YY}(y_j),\, S_{YX}(y_j)\big) \tag{4}$$

with $\mathrm{SD}(X,Y) = \sqrt{\mathrm{SD}^2(X,Y)}$.

When X and Y are drawn from the same distribution, each point's intra- and cross-distribution distances have identical sorted profiles, yielding $\mathrm{SD} = 0$. When the distributions differ in density, shape, or location, the sorted distance profiles diverge.

Theoretical Properties. *Identity of indiscernibles.* By convention, self-distances (the zero diagonal) are masked and

excluded. When evaluating an identical set $\mathrm{SD}(X, X)$, the intra- and cross-signatures for every point are identical, so $\mathrm{SD}(X, X) = 0$.

Metric properties and triangle inequality. SD satisfies non-negativity ($\mathrm{SD}(X, Y) \geq 0$), identity of indiscernibles ($\mathrm{SD}(X, Y) = 0 \Leftrightarrow X = Y$), and symmetry by construction. As shown by Székely and Rizzo (2013), satisfying the triangle inequality unconditionally requires the underlying kernel to be of conditionally negative definite type (see Appendix C). Establishing this property analytically for the sorted signature kernel remains an open problem. All simulations in this work are consistent with metric behavior under the square-root formulation $\mathrm{SD} = \sqrt{\mathrm{SD}^2}$.

Upper and lower bounds (Appendix A, Theorem 1). The asymmetric components and their symmetric average SD^2 are bounded below by half the ED and above by the exact 1-Wasserstein distance:

$$\frac{1}{2}\mathcal{E}(X, Y) \;\leq\; \mathrm{SD}^2(X, Y) \;\leq\; W_1(X, Y) \;\leq\; \mathbb{E}\|X - Y\| \qquad (5)$$

The upper bound follows because the signature function $f_z(u) = \|z - u\|$ is 1-Lipschitz, so W_1 is non-increasing under the pushforward; the lower bound follows from the mean bound on 1D Wasserstein distances and Jensen's inequality. SD^2 is an $\mathcal{O}(n^2)$ statistic bounded above by the exact W_1 distance.

Scale-invariant normalization (SD_{norm}). The rightmost bound in the chain, $\mathbb{E}\|X - Y\|$, provides a scale factor that normalises SD^2 into a unit interval. This permits the definition of an absolute, scale-invariant index:

$$\mathrm{SD}_{\mathrm{norm}}(X, Y) = \sqrt{\frac{\mathrm{SD}^2(X, Y)}{\mathbb{E}\|X - Y\|}} \in [0, 1] \qquad (6)$$

Computational complexity. Pairwise distances scale as $\mathcal{O}((n + m)^2)$, and the per-point sorting scales as $\mathcal{O}((n + m) \cdot \max(n, m) \log \max(n, m))$. SD thus has the same complexity class as ED.

Asymmetric decomposition and computational efficiency. Equation 4 decomposes naturally into two directed components. In a generative setting where X denotes generated samples and Y real reference data, these components admit a precision/recall interpretation (Kynkäänniemi et al., 2019):

$$\mathrm{SDiv}(X \parallel Y) = \frac{1}{n}\sum_{i=1}^{n} \delta\big(S_{XX}(x_i),\, S_{XY}(x_i)\big) \quad \text{(Precision / Fidelity)}$$

$$\mathrm{SDiv}(Y \parallel X) = \frac{1}{m}\sum_{j=1}^{m} \delta\big(S_{YY}(y_j),\, S_{YX}(y_j)\big) \quad \text{(Recall / Coverage)}$$

so that $\mathrm{SD}^2(X, Y) = \frac{1}{2}\mathrm{SDiv}(X \parallel Y) + \frac{1}{2}\mathrm{SDiv}(Y \parallel X)$. $\mathrm{SDiv}(X \parallel Y)$ (precision) evaluates whether each generated sample's neighborhood structure matches that of the real manifold; $\mathrm{SDiv}(Y \parallel X)$ (recall) evaluates whether each real sample's local neighborhood is faithfully reproduced by the generated distribution. This decomposition has a practical consequence for iterative algorithms: when Y is a fixed reference, the intra-reference signatures S_{YY} can be precomputed once and cached, reducing per-step cost from $\mathcal{O}(n^2 + nm)$ to $\mathcal{O}(nm)$.

Finite-sample inference and hypothesis testing. SD supports a standard permutation-based two-sample test. Given samples X (size n) and Y (size m), compute the observed statistic $d_{\mathrm{obs}} = \mathrm{SD}(X, Y)$. Pool both sets into $Z = X \cup Y$ and, for $b = 1, \ldots, B$ permutations, randomly partition Z into groups of size n and m and compute $d_b = \mathrm{SD}(Z_{1:n}, Z_{n+1:n+m})$. The empirical p-value is $p = (\sum_{b=1}^{B} \mathbb{1}[d_b \geq d_{\mathrm{obs}}] + 1)/(B + 1)$, following the pseudocount convention of Phipson and Smyth (2010) to ensure strictly positive estimates and controlled Type I error. The test reuses the precomputed full pairwise distance matrix, permuting indices rather than recomputing distances, reducing the cost of B permutations from $\mathcal{O}(B \cdot n^2)$ to $\mathcal{O}(B \cdot n \log n)$.

This permutation test evaluates Fisher's conditional null: whether the observed partition produces a distance larger than expected under random label assignment, conditional on the specific sample at hand (Ernst, 2004). Its resolution is bounded by the combinatorial space $\binom{n+m}{n}$, which for moderate samples ($n = m = 15$, $\binom{30}{15} \approx 1.5 \times 10^8$) can yield mathematically valid p-values well below 0.001. Such p-values reflect exact label-assignment probability, but do not account for the sampling variance inherent in drawing a finite cohort from a broader population. To complement the randomisation test with a measure of population-level stability, we provide bootstrap confidence intervals (Efron and Tibshirani, 1993): resampling the observed groups with replacement over B_{boot} iterations and recomputing SD yields an empirical estimate of the distance's sampling distribution, whose 95% interval quantifies how much the structural difference might vary across hypothetical replicate cohorts.

Column Distance and density-level constraints

The signature matrix $S_{XY} \in \mathbb{R}^{N \times K}$ can be integrated along either axis, yielding complementary metrics. SD integrates *row-wise*, comparing per-point neighbourhood topology (Eq. 4). A natural complement integrates *column-wise*: for each rank k, the population distribution of k-th nearest-neighbour distances should match across the two sets. We define the asymmetric Column Divergence and the **Squared Column Distance (CD^2)**:

$$\mathrm{CDiv}(X \parallel Y) = \frac{1}{K}\sum_{k=1}^{K} W_1\big(\{S_{XX}[\cdot, k]\},\ \{S_{XY}[\cdot, k]\}\big)$$

$$\mathrm{CD}^2(X, Y) = \tfrac{1}{2}\,\mathrm{CDiv}(X \parallel Y) + \tfrac{1}{2}\,\mathrm{CDiv}(Y \parallel X) \qquad (7)$$

Because each column is one-dimensional, W_1 reduces to sorting, with $\mathrm{CD} = \sqrt{\mathrm{CD}^2}$. CD constrains the *global density level-set structure* (the fraction of points within each nearest-neighbour shell) while SD constrains the local topology. Combining them via their Pythagorean (L2 product-space) combination preserves the conditional negative definiteness of the individual squared terms, ensuring Hilbert-space embeddability via Schoenberg's theorem:

$$\mathrm{CSD}(X, Y) = \sqrt{\mathrm{SD}^2(X, Y) + \mathrm{CD}^2(X, Y)} \qquad (8)$$

The equal weighting corresponds to the L2 product-space norm of the row and column signals; task-dependent weightings are a natural extension but are not explored here. However, column sorting marginalises spatial identity: CD does not track *which* points contribute to each density level.

Grounded Signature Distance

SD compares each point's neighbourhood topology to the population average but does not enforce spatial correspondence with any specific reference point. CD addresses this through population-level density matching but marginalises point identity.

Grounded Signature Distance (GSD) provides a more direct solution by grounding each point to its nearest reference neighbour. We define the asymmetric Grounding Divergence and the symmetric Squared Grounded Signature Distance:

$$\mathrm{GDiv}(X \parallel Y) = \frac{1}{n}\sum_{i=1}^{n} \delta\big(S_{XY}(x_i),\, S_{YY}(y_{\mathrm{nn}_{X\to Y}(i)})\big)$$

$$\mathrm{GSD}^2(X,Y) = \mathrm{SD}^2(X,Y) + \tfrac{1}{2}\,\mathrm{GDiv}(X \parallel Y) + \tfrac{1}{2}\,\mathrm{GDiv}(Y \parallel X) \tag{9}$$

where $y_{\mathrm{nn}_{X\to Y}(i)}$ is the nearest neighbour in Y to $x_i \in X$, $\delta(\cdot,\cdot)$ is the pointwise signature divergence (Eq. 3), and $\mathrm{GSD}(X,Y) = \sqrt{\mathrm{GSD}^2(X,Y)}$. The grounding term evaluates whether each point's distance profile matches that of its nearest neighbour in the opposing set, spatially grounding the profile comparison. This captures density-level information that CD addresses through column sorting, with the NN lookup and profile comparison reusing the already-computed distance and signature matrices. When used as a training loss, the NN targets are treated as fixed reference values (detached from the computational graph), providing a stable spatial anchor without amplifying gradient noise.

Results

Controlled experiments use synthetic two-dimensional distributions; all biological experiments use The Cancer Genome Atlas (TCGA) pan-cancer gene expression dataset (The Cancer Genome Atlas Research Network, 2008).

Sensitivity to distributional differences

To illustrate the behaviour of SD, CSD, and ED under controlled distributional perturbations, we consider a two-dimensional Gaussian scenario (Figure 2). A reference distribution X of 200 points from a standard two-dimensional Gaussian is tested against three alternatives: identical, location shift, and density change.

All three metrics correctly identified the identical scenario as non-significant and the location shift as significant. In the density-change scenario, SD and CSD detect the perturbation while ED does not.

The null distribution, constructed from random splits of the pooled data, provides the baseline for interpreting these results. For a uniform contraction, the barycenter is preserved and the mean cross-distance shifts only modestly; the perturbation is therefore effectively invisible to ED, which relies entirely on mean pairwise distances. SD detects it because the contraction tightens the intra-distribution signature of the contracted set relative to the cross-signature, generating a systematic shape difference in the sorted arrays that the mean operation in ED cannot resolve.

Per-point loss landscapes

Beyond aggregate distance computation, the un-averaged per-point contributions to SD reveal its behavior as a spatially resolved generative objective. For each candidate point g generated against a true data manifold Y, we define the pointwise generative penalties $\delta(g)$. For Energy Distance, this evaluates the expected physical distance to the manifold: $\delta_{\mathrm{ED}}(g) = \mathbb{E}_{y\sim Y}\|g-y\|$. For Signature Distance, this evaluates the topological divergence of g's local neighborhood: $\delta_{\mathrm{SD}}(g) = W_1(S_{gY}, \bar{S}_{YY})$, where $S_{gY} = \mathrm{sort}(\|g-Y\|)$ is the cross-signature from g to Y, and $\bar{S}_{YY}$ is the expected intra-manifold signature.

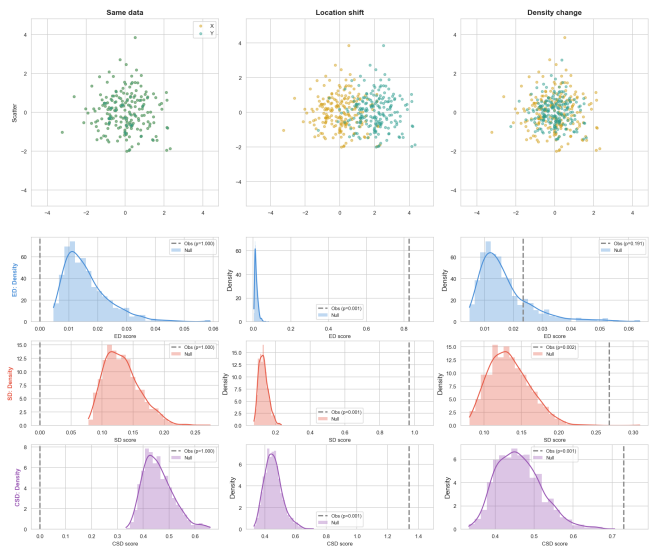

Fig. 2: Sensitivity comparison of SD, CSD, and ED on controlled two-dimensional scenarios. *Top row:* scatter plots of X (orange) and Y (blue). *Rows 2–4:* permutation-test null distributions (shaded) and observed values (dashed lines) for ED, SD, and CSD. *Left (same distribution):* X and Y are drawn from the same Gaussian; all metrics correctly fall within their null. *Centre (location shift):* Y is shifted; all three detect the difference ($p = 0.001$). *Right (density change):* $Y = X \times 0.7$ (uniform contraction preserving shape). This is the critical test: ED is non-significant ($p = 0.19$) because the mean distance is nearly unchanged, while both SD ($p = 0.002$) and CSD ($p = 0.001$) detect the contraction through changes in the *shape* of the distance profile, not just its mean.

We evaluated these landscapes on two topologies that highlight the limitations of ED as a generative objective: (a) two separated clusters and (b) a continuous ring (Figure 3).

In the two-cluster scenario, a point g placed in the equidistant gap satisfies $\mathbb{E}\|g-Y\| \approx \mathbb{E}_{y\sim Y}\|y-Y\|$: the mean physical distance to the opposing cloud matches that of a genuine in-distribution point, so ED assigns the same penalty to the interpolation artifact as to real data, resulting in an uninformative landscape in the gap region. By contrast, g's signature S_{gY} is strongly bimodal, departing markedly from the unimodal manifold signature $\bar{S}_{YY}$, so $\delta_{\mathrm{SD}}(g) \gg 0$: the artifact receives a large and distinguishing penalty.

The ring topology reveals a structural mismatch between ED's global minimum and the true data manifold: the geometric center of the ring uniquely minimizes $\mathbb{E}\|g-Y\|$, because it equalizes physical distances to all arc positions. The ED minimum therefore lies in the empty interior. The SD penalty at the center is instead large: S_{gY} for a central point is nearly uniform across all ring distances, while the true intra-ring signature $\bar{S}_{YY}$ is concentrated near zero (within-ring neighbor distances). The SD minimum aligns with the ring perimeter, consistent with the contour plots in Figure 3.

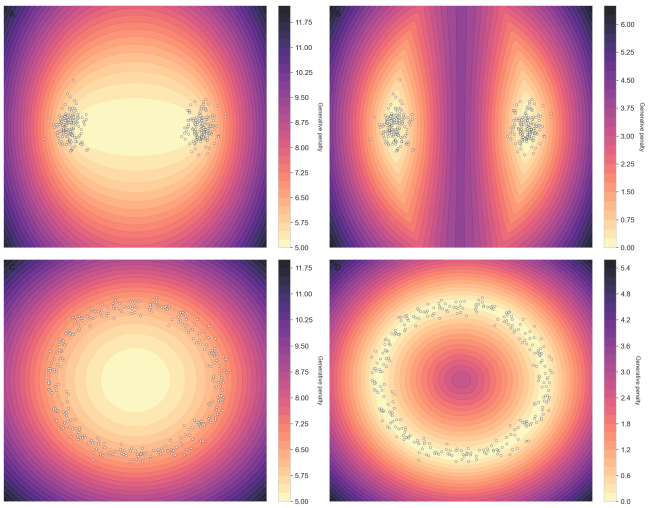

Fig. 3: Per-point generative loss landscapes reveal how a single new point would be penalised at each location. *Top row (two clusters):* the ED landscape (left) is nearly flat in the inter-cluster gap, offering no gradient signal to guide a generated point away from the empty region; SD (right) produces a pronounced penalty ridge between the clusters, showing that a point placed there would have an anomalous distance profile relative to both cluster distributions. *Bottom row (ring):* the ED minimum is at the ring's empty centre (where mean distance to all reference points is minimised), while the SD minimum traces the ring perimeter itself. This explains why ED-trained generators tend to place mass at barycenters and in gaps, while SD-trained generators follow the support geometry.

Interpolation sensitivity

Two TCGA sub-populations (A and B, 400 points each) were used to create interpolated points: $A_\alpha = (1-\alpha)\cdot A_{\text{matched}} + \alpha \cdot B_{\text{matched}}$, where matching was performed via optimal assignment (Kuhn, 1955). $\text{SD}(A_\alpha, B)$, $\text{CSD}(A_\alpha, B)$, and $\text{ED}(A_\alpha, B)$ were measured as α increased from 0 to 1. To contextualize the magnitudes, the null mean of each metric was computed from random TCGA subsets.

ED decreased monotonically toward zero, while SD and CSD exhibited non-monotonic structural penalties, peaking near $\alpha = 0.5$ (Figure 4). At $\alpha = 0.5$, ED assessed the interpolated set as physically closer to B than a random real subset (ED/null $= 0.76$), while SD assessed it as structurally off-manifold (SD/null $= 1.37$). CSD showed the strongest penalty (CSD/null $= 2.22$), combining the row-level signature mismatch of SD with the column-level density mismatch of CD.

This divergence has a direct geometric explanation in terms of the signature shape argument developed in Section 3.2. Linearly interpolated points sit in sparse, off-manifold regions between A and B. Their cross-signatures S_{gB} are structurally bimodal, strongly resembling those of the two-cluster interpolation artifact, because nearby neighbours from B are far closer than distant neighbours, creating a pronounced gap in the sorted array. The intra-manifold signature $\bar{S}_{BB}$ of true B points, by contrast, is unimodal and compact. SD registers this signature mismatch; ED, which collapses the entire distance profile to its mean, does not. As $\alpha \to 1$, interpolated points converge onto B, their signatures become unimodal, and both metrics approach zero together.

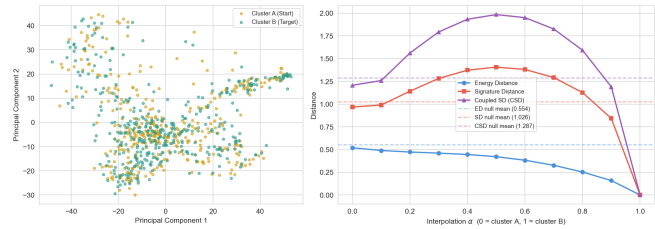

Fig. 4: Interpolation sensitivity on TCGA transcriptomic data, testing whether distributional metrics can detect synthetic samples that lie "between" real populations. X is a 400-sample sub-population; Y is a matched sub-population. Interpolated sets $A_\alpha = (1-\alpha)A + \alpha B$ are constructed for $\alpha \in [0, 1]$; at $\alpha = 0$, the interpolated set equals A itself. Dashed lines show the null mean from random TCGA subsets. ED (left) decreases monotonically as the interpolated set physically converges toward B, and never exceeds the null; it cannot detect that interpolated biological samples occupy unphysical regions of gene expression space. SD (centre) peaks at $\alpha \approx 0.5$ and CSD (right) shows an even stronger non-monotonic penalty, reflecting that the sorted distance profiles of interpolated points are systematically anomalous even when their mean distances are not.

Data expansion via guided Markovian sampling

Because SD is differentiable via automatic differentiation through `torch.sort`, the gradient $\nabla_x \text{SD}(\text{pop}, A_k)$ is computable for any candidate position x, making SD (and its extensions) usable directly as the potential energy of a Langevin diffusion (Welling and Teh, 2011). This requires no generative model: the expansion operates as gradient descent in the space of empirical distributions, regularised by Cholesky-preconditioned Gaussian noise that confines displacement to directions of high variance in the seed set.

Using TCGA data clustered via Leiden (Traag et al., 2019), the largest cluster was partitioned into seed A, validation B, and hidden test C. An initial population was constructed from A using K-means (Lloyd, 1982) centroids scaled to match $|A|$. The population was then evolved for 500 steps under Langevin dynamics:

$$x_{t+1} = \text{clip}\Big(x_t - \eta\, \nabla_x\, \mathcal{D}(x_t,\, A_k) + \Sigma_{A_k}^{\frac{1}{2}} \varepsilon_t,\; 0\Big) \tag{10}$$

where $\varepsilon_t \sim \mathcal{N}(0, I)$, Σ_{A_k} is the empirical covariance of the seed set (Cholesky-normalised), and $\mathcal{D} \in \{\text{ED, SD, CSD}\}$ is the guidance loss. Critically, the gradient is taken with respect to A_k only: B is deliberately excluded from the training signal to preserve it as an uncontaminated validation reference.

The chain is stopped at the epoch that minimises the following stopping score:

$$\mathcal{S}(\text{pop}) = \big(\mathcal{D}(\text{pop}, A_k) - \mathcal{D}(\text{pop}, B)\big)^2 + \big(\mathcal{D}(\text{pop}, A_k) - \mathcal{D}(A_k, B)\big)^2 \tag{11}$$

This objective is satisfied when $\mathcal{D}(\text{pop}, A_k) \approx \mathcal{D}(\text{pop}, B) \approx \mathcal{D}(A_k, B)$: the expanded population has become equidistant from A and B in the chosen metric, reaching the distributional midpoint between the two folds.

To stabilise the stopping epoch, we applied a bootstrap protocol: $K = 5$ replicates draw A_k with replacement from A,

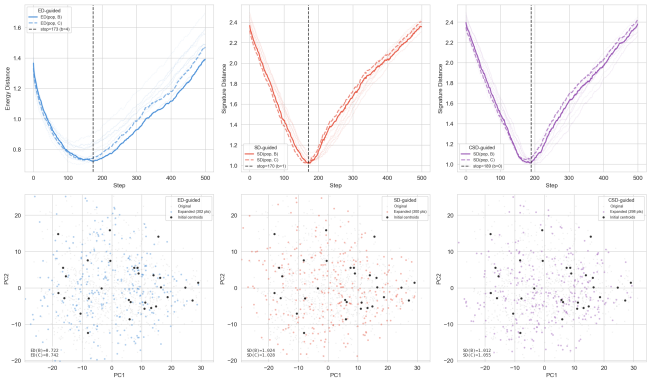

Fig. 5: Langevin expansion on the largest TCGA Leiden cluster, demonstrating SD as a potential energy for model-free data augmentation. New samples are generated by gradient descent on the distributional distance to the cluster, then stopped when the held-out validation distance $\mathcal{S}$ (Eq. 11) reaches its minimum. *Top row:* distance traces across 500 steps for the best of $K = 5$ bootstrap replicates; solid line = distance to validation set B, dashed = distance to held-out C; vertical dashed line marks the stopping epoch. All three metrics can guide expansion, but SD-guided samples (centre) most closely track the held-out population, suggesting better generalisation beyond the training cluster. *Bottom row:* PCA projections at the stopping epoch; grey = original cluster, coloured = expanded samples. Metrics annotated bottom-left.

each producing an independent chain with its own covariance Σ_{A_k}. The replicate with the lowest $\mathcal{D}(\text{pop}, B)$ at its optimal step is selected; C is used exclusively for final evaluation and never influences stopping.

Both SD- and ED-guided traces show clear U-shaped minima within 500 steps, as does CSD-guided expansion (Figure 5). The SD-guided stopping epochs are stable across bootstraps (range $\approx$ 30 steps), while the ED-guided epoch shows higher variance ($\approx$ 70 steps), consistent with SD's greater structural sensitivity.

The stopping criterion (Equation 11) encodes a soft indistinguishability constraint: the expanded population should be as similar to B as the seed A_k is, under the chosen metric. The resulting U-shaped stopping score provides a principled, observable criterion whose minimum can be identified from the validation fold alone. The held-out fold C is used exclusively to verify that the stopping generalises; the near-zero gap between $\mathcal{D}(\text{pop}, B)$ and $\mathcal{D}(\text{pop}, C)$ at the optimal step confirms that the procedure does not overfit the stopping epoch.

Because the expansion uses only a local reference set to anchor the generated points in feature space, it can be applied independently to each subpopulation in a dataset. Partitioning the data and expanding each partition separately allows the full manifold to be densely represented at arbitrary scale, with the reference serving only to pin each expansion to the correct region of the space.

SD as differentiable generative loss

Because the quantile sorting operations underlying the signatures are tracked by automatic differentiation frameworks (e.g., `torch.sort`), SD can be used as a differentiable loss function

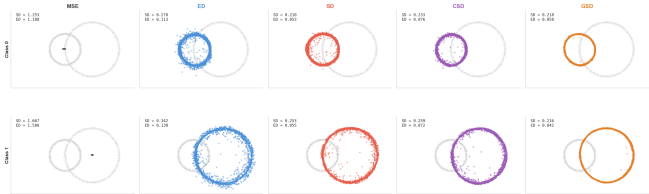

Fig. 6: Generative loss comparison on two intertwining circles of different radii, a topology that cannot be recovered by a mean-seeking loss. Each column shows a different loss function; rows show class 0 (inner circle) and class 1 (outer circle). MSE collapses both classes to their respective barycenters, producing point masses. All four distributional losses recover the circular structure, confirming that distributional objectives can learn topology from an unpaired generation task where MSE cannot.

for training generative neural networks (implementation details in Appendix B). We first confirm gradient stability on controlled toy problems, then demonstrate the approach at scale on a tissue-conditioned generation task using TCGA gene expression data.

A class-conditioned MLP generator (five hidden layers with ReLU activations) was trained to map isotropic Gaussian noise to a target distribution of two intertwining circles of different radii (Figure 6). All models were evaluated by SD and ED score (lower is better) on 2,000 generated versus 2,000 real points per class.

MSE collapses to the class-conditional barycenter, as expected when a pointwise loss is applied to an unpaired generation task. The distributional losses all recover the circular structure. We additionally tested network normalization and loss decomposition on the same task (Figure 10, Appendix B). Batch normalization is incompatible with distributional losses: normalizing activations across the batch cancels the inter-sample variance signal that these losses depend on, causing fragmentation of the generated distribution.

Heteroscedastic sinusoidal regression. To illustrate that these properties extend from topology-learning to continuous conditional regression, we trained the same noise-conditioned MLP on a heteroscedastic sinusoidal task: $y = \sin(t) + \sigma(t) \cdot \xi$, with narrow noise ($\sigma = 0.06$) for $t \in [0, \pi]$ and wide noise ($\sigma = 0.3$) for $t \in [\pi, 2\pi]$ (Figure 7). The noise vector ε in the model input is never paired with the specific ξ used to draw a target, so a point-prediction loss has little incentive to exploit it; distributional losses, by contrast, naturally use ε to span the output distribution.

All distributional losses correctly capture the heteroscedastic structure. GSD produces the tightest fit to the ground-truth envelope, consistent with the spatial grounding provided by the nearest-neighbour cross term.

Tissue-conditioned generation on TCGA landmark genes. Having confirmed gradient stability on controlled tasks, we applied the losses to a biologically relevant generation problem designed to expose how each loss function handles high-dimensional, multi-population structure under a deliberate capacity constraint. TCGA provides RNA expression measurements across 978 landmark genes (Subramanian et al., 2017) and 24 tissue types. A tissue-conditioned generator was trained to map a low-dimensional noise vector ($\varepsilon \in \mathbb{R}^{32}$) concatenated with a tissue-type label to a 978-dimensional gene expression profile. The 32-dimensional

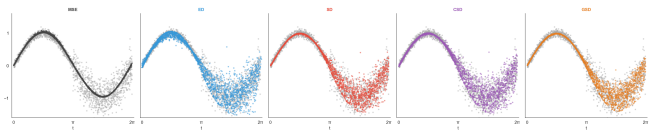

Fig. 7: Heteroscedastic sinusoidal regression ($y = \sin(t) + \sigma(t)\xi$, narrow noise for $t \in [0, \pi]$, wide noise for $t \in [\pi, 2\pi]$). The noise input ε is never paired with the specific ξ used to draw the target, so the model must learn to use ε to span the output distribution. MSE collapses to the conditional mean, discarding the noise input entirely. All distributional losses recover the heteroscedastic spread, matching both the narrow and wide noise bands. GSD produces the tightest fit to the ground-truth envelope, consistent with the spatial grounding provided by its nearest-neighbour coupling term.

input is a deliberate bottleneck: with 978 output dimensions, a higher-capacity input would reduce the performance gap between losses, masking their technical differences. The constrained regime forces each loss to prioritise different aspects of the distributional geometry, revealing their distinct inductive biases.

Each loss was trained with a *glocal* protocol, combining a global term (the distance computed on the full mixed-tissue batch) with local per-tissue terms (the mean of per-tissue distances within each tissue group in the batch):

$$\mathcal{L}_{\text{glocal}} = \frac{1}{2}\left[\mathcal{D}(G_{\text{all}}, R_{\text{all}}) + \frac{1}{|\mathcal{T}_b|}\sum_{t \in \mathcal{T}_b} \mathcal{D}(G_t, R_t)\right] \tag{12}$$

where $\mathcal{T}_b$ is the set of tissue types present in the batch, and G_t, R_t are the generated and real samples for tissue t. The global term preserves cross-tissue manifold structure; the local terms ensure within-tissue distributional fidelity. All losses are trained with the same protocol.

The quality of each generator was evaluated on a held-out test set of 2,032 samples via two independent criteria: (1) a downstream tissue-type classifier trained exclusively on 10× synthetic data and evaluated on real test samples (accuracy, macro F1), measuring whether the synthetic data captures tissue-discriminative structure; and (2) a 1-NN distributional analysis measuring coverage (fraction of real training points that are nearest-neighbour to at least one synthetic point), normalised entropy of the NN-count distribution, and mean NN distance.

Table 1. Tissue-conditioned generation on TCGA landmark genes (978 genes, 24 tissues, noise dim = 32). All losses are trained with the same glocal protocol (Equation 12), ensuring fair comparison; only the distance metric $\mathcal{D}$ varies. Values in bold indicate the best result per column.

Loss	Acc (%)	F1 (%)	Coverage	Entropy	NN-dist
MSE	83.1	78.2	0.004	0.329	**12.8**
ED	84.3	80.5	0.173	0.663	25.9
SD	86.1	83.2	0.230	0.689	24.8
CSD	88.3	84.7	0.257	0.704	24.4
GSD	**89.9**	**84.7**	**0.268**	**0.736**	**21.0**

The glocal protocol reveals a clear hierarchy among distributional losses. All losses achieve comparable classification accuracy (83-90%), including MSE, because the per-tissue local term provides within-tissue paired supervision even for a pointwise loss. Among distributional losses, GSD leads on every metric: highest accuracy, coverage, entropy, and lowest NN distance among distributional losses. The symmetric nearest-neighbour grounding both pulls generated points closer to the real manifold and encourages broader coverage, since each real point's neighbourhood is explicitly targeted from both directions. CSD provides the next-best results, with its column-level density constraints achieving strong coverage and accuracy. Per-tissue F1 scores (Figure 11, Appendix B) confirm that these aggregate differences are consistent across tissues.

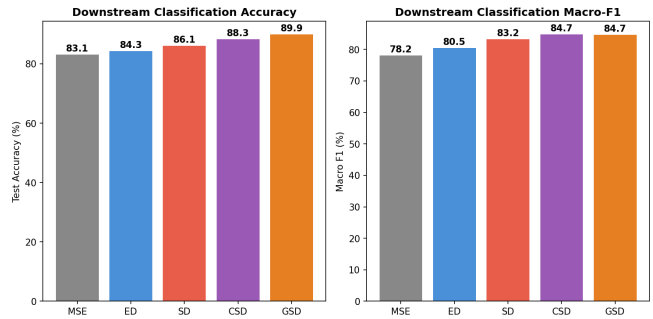

Fig. 8: Downstream tissue classification on real test data, using classifiers trained exclusively on synthetic data. Higher accuracy indicates that the synthetic data captures tissue-discriminative gene expression patterns. The glocal protocol allows even MSE to achieve reasonable classification (83%), but distributional losses consistently outperform it. GSD leads with 89.9%, consistent with the spatial grounding provided by its nearest-neighbour coupling term.

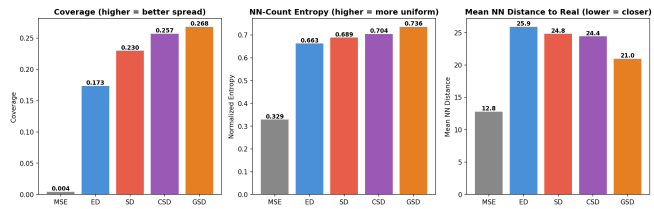

Fig. 9: 1-NN distributional analysis quantifying how well synthetic data covers the real manifold. *Coverage* (left): the fraction of real training points that are the nearest neighbour to at least one synthetic point (higher means the generator reaches more of the data manifold). *Entropy* (centre): uniformity of the NN-count distribution (higher means generated mass is spread evenly rather than clumped). *NN distance* (right): mean distance from synthetic to nearest real point (lower means closer proximity). MSE achieves the lowest NN distance (its barycentric outputs are near each tissue's mean) but near-zero coverage (0.4%), demonstrating that proximity alone does not imply distributional fidelity. GSD leads on all distributional metrics: highest coverage, entropy, and lowest NN distance among distributional losses, reflecting how spatial grounding both pulls generated points closer to the real manifold and encourages broader coverage.

Discussion

Interpretation and broader implications

The sensitivity and loss landscape experiments share a common geometric explanation. Energy distance collapses the full distance profile from a point to the opposing distribution into its mean, discarding all shape information. Signature Distance retains the full sorted profile, making it sensitive to any distributional property encoded in the shape of the radial distance distribution: variance, skewness, and multi-modality. The advantage of SD over ED is therefore determined by whether the distributional difference in question alters the shape of these profiles. The signature matrix encodes complementary information along its two axes: CD captures population-level density-scale constraints by sorting each column independently, while SD captures per-point topology via rows. Their Pythagorean combination (CSD) adds within-population density constraints, while the nearest-neighbour grounding in GSD addresses this more directly, preserving point identity at the same asymptotic cost.

A practical consequence for generative model evaluation in biology (Lopez et al., 2018; Marouf et al., 2020) is that models suffering from interpolation artifacts (Goodfellow et al., 2014; Kingma and Welling, 2014; Rezende and Mohamed, 2015) can produce synthetic samples that receive favorable scores under barycenter-based evaluators like FID or MMD despite lacking realistic local structure. SD provides a more sensitive criterion for detecting such artifacts because it evaluates the full distance profile shape rather than its mean.

The model-free character of the Langevin expansion is particularly relevant for biological data augmentation: no generative architecture is required, the noise covariance is estimated directly from the seed cluster, and the objective is the same distributional metric used for evaluation. Unlike interpolation-based augmentation methods (Chawla et al., 2002), which produce off-manifold artifacts detectable by SD (Section 3.3), the Langevin approach generates samples that respect the distributional geometry of the seed population. This makes it immediately applicable to any modality for which a pairwise distance can be computed, such as protein structures, chemical fingerprints or multi-omic profiles (Argelaguet et al., 2018), without any retraining. It is especially relevant in settings where foundation model embeddings are available but task-specific labels are not. In single-cell biology, for example, pretrained encoders (Theodoris et al., 2023; Cui et al., 2024) provide rich representations of cellular state, yet labelled examples of a specific perturbation effect (Dixit et al., 2016) may comprise only a handful of observations. Distributional losses allow new samples to be generated directly in the embedding space, guided solely by the geometry of the observed population, without requiring paired supervision or finetuning of the foundation model.

Distributional losses address a different problem from pointwise regression: they match *distributions* where no pointwise correspondence exists. They serve as natural objectives for unconditional generation, data augmentation, and model-free sampling, and may complement pointwise losses as hybrid objectives or as alternatives to adversarial discriminators where mode coverage guarantees are desirable. The glocal protocol (Equation 12) demonstrates that batch composition is essential for unlocking their full potential on multi-population data, but the design space of batch strategies remains largely unexplored.

Limitations and future directions

SD carries an additional $\mathcal{O}(n \log n)$ factor per point from the sorting step beyond the $\mathcal{O}(n^2)$ pairwise distance computation shared with ED. For the dataset sizes considered here ($n < 10{,}000$) this is negligible; at larger scales, subsampling or approximate nearest-neighbor methods can reduce this cost without qualitatively changing the metric behavior.

A formal proof that the sorted signature kernel is of conditionally negative definite type remains an open problem (Appendix C). The hard nearest-neighbour assignment in GSD is a simple, parameter-free approximation to optimal transport coupling in signature space; smoother alternatives, such as softmax-weighted profile matching or Sinkhorn-based optimal plans (Cuturi, 2013) over signature rows, may improve calibration at the cost of additional computation.

SD is defined for any metric space, requiring only pairwise distances, making it applicable to protein structure comparison (RMSD-based distances), chemical similarity (Tanimoto distances), and multi-omic integration. The distance signature shares underpinnings with the Vietoris–Rips filtration in computational topology (Edelsbrunner and Harer, 2010): a point's signature encodes its birth times in a Vietoris–Rips complex. Formalizing connections to persistent homology may yield further theoretical insights.

Concluding remarks

We introduced Signature Distance, a distributional comparison method that generalises energy distance by retaining the full sorted pointwise distance profile rather than collapsing it to a scalar mean. SD matches the $\mathcal{O}(n^2)$ computational cost of ED while capturing local density structure, manifold curvature, and off-manifold interpolation artifacts that scalar reductions cannot resolve. Extensions along complementary axes of the signature matrix (Column Distance for population-level density constraints, and nearest-neighbour grounding (GSD) for spatial correspondence) address different aspects of distributional fidelity.

On TCGA gene expression data, SD and its extensions translate these geometric properties into practical advantages: as two-sample tests, as differentiable generative training objectives with a glocal protocol for multi-population data, and as direct potential energies for model-free Langevin data expansion. The tissue-conditioned generation experiment demonstrates that under capacity constraints, distributional losses extract substantially more useful structure per bit of latent capacity than pointwise regression, and that batch composition is a key design consideration when deploying these losses at scale. Code and data to reproduce all experiments are available on GitHub.

Data Availability

The TCGA data used in this work was downloaded from the GDC Data Portal (`https://portal.gdc.cancer.gov/`). The code used to perform the experiments and reproduce the results in this paper is available on GitHub at: `https://github.com/lazzaronico/signature-distance`.

Acknowledgments and Funding

The results shown here are in whole or part based upon data generated by the TCGA Research Network[1]. This work is partially funded by the EU through the 3DSecret project under the HORIZON-EIC-2022-PATHFINDER-OPEN-01-01 programme (grant no. 101099066).

Appendix A: Upper and Lower Bounds for Squared Signature Distance

Theorem 1. *Let μ and ν be probability measures on $\mathbb{R}^d$. Define the population-level asymmetric divergence*

$$\mathrm{SDiv}(\mu \parallel \nu) = \int W_1\big(S_\mu(z),\, S_\nu(z)\big)\, d\mu(z)$$

where $S_\mu(z)$ denotes the distribution of radial distances from z under μ (i.e. the pushforward $(f_z)\#\mu$ with $f_z(u) = \|z-u\|$), and the symmetric squared distance $\mathrm{SD}^2(\mu,\nu) = \frac{1}{2}\mathrm{SDiv}(\mu \parallel \nu) + \frac{1}{2}\mathrm{SDiv}(\nu \parallel \mu)$. Then:

$$\frac{1}{2}\mathcal{E}(\mu,\nu) \;\le\; \mathrm{SD}^2(\mu,\nu) \;\le\; W_1(\mu,\nu) \;\le\; \mathbb{E}_{x\sim\mu,\, y\sim\nu}\|x-y\| \quad (13)$$

Proof of the upper bound ($\mathrm{SD}^2 \le W_1$). For any reference point $z \in \mathbb{R}^d$, the signature $S_\mu(z) = (f_z)\#\mu$ is the pushforward of μ through $f_z(u) = \|z-u\|$. By the reverse triangle inequality, f_z is 1-Lipschitz: $|f_z(u) - f_z(v)| = \big|\|z-u\| - \|z-v\|\big| \le \|u-v\|$. A standard property of optimal transport states that the 1-Wasserstein distance is non-increasing under 1-Lipschitz pushforwards: any coupling γ for (μ,ν) induces a coupling for $(S_\mu(z),\, S_\nu(z))$ with cost $\int |f_z(u)-f_z(v)|\, d\gamma \le \int \|u-v\|\, d\gamma$; taking the infimum gives

$$W_1\big(S_\mu(z),\, S_\nu(z)\big) \le W_1(\mu,\nu) \quad \text{for every } z.$$

Taking the expectation over $z \sim \mu$ establishes $\mathrm{SDiv}(\mu \parallel \nu) \le W_1(\mu,\nu)$; by symmetry, $\mathrm{SDiv}(\nu \parallel \mu) \le W_1(\mu,\nu)$; their average SD^2 inherits the bound. Finally, $W_1(\mu,\nu) \le \mathbb{E}\|x-y\|$ because the independent coupling is a valid (non-optimal) transport plan.

Proof of the lower bound ($\mathrm{SD}^2 \ge \frac{1}{2}\mathcal{E}$). For any two one-dimensional distributions A and B, the 1-Wasserstein distance satisfies the mean bound: $W_1(A,B) \ge |\mathbb{E}[A] - \mathbb{E}[B]|$ (the identity function is 1-Lipschitz, so this follows from the Kantorovich–Rubinstein duality with test function $f(t) = t$). Applying this to the recall term at reference $y \sim \nu$:

$$W_1\big(S_\nu(y),\, S_\mu(y)\big) \ge \Big|\mathbb{E}_{y'\sim\nu}\|y-y'\| - \mathbb{E}_{x\sim\mu}\|y-x\|\Big|$$

Averaging over $y \sim \nu$ and applying Jensen's inequality ($\mathbb{E}|h| \ge |\mathbb{E}[h]|$):

$$\mathrm{SDiv}(\nu \parallel \mu) \ge \Big|\mathbb{E}_{x,y}\|x-y\| - \mathbb{E}_{y,y'}\|y-y'\|\Big| \ge \mathbb{E}\|X-Y\| - \mathbb{E}\|Y-Y'\|$$

where the last step uses $|a-b| \ge a-b$. By exact symmetry, $\mathrm{SDiv}(\mu \parallel \nu) \ge \mathbb{E}\|X-Y\| - \mathbb{E}\|X-X'\|$. Summing the two lower bounds and collecting:

$$\begin{aligned}\mathrm{SD}^2(\mu,\nu) &= \tfrac{1}{2}\mathrm{SDiv}(\mu \parallel \nu) + \tfrac{1}{2}\mathrm{SDiv}(\nu \parallel \mu)\\ &\ge \tfrac{1}{2}\Big(2\mathbb{E}\|X-Y\| - \mathbb{E}\|X-X'\| - \mathbb{E}\|Y-Y'\|\Big) = \tfrac{1}{2}\mathcal{E}(\mu,\nu)\end{aligned}$$

■

[1] `https://www.cancer.gov/tcga`

Appendix B: Signature Distance as a Differentiable Generative Loss

Because autograd tracks indexing operations during `torch.sort`, SD^2 is piecewise differentiable and can be used as a structural loss function for deep generative modeling. The following implementation choices are recommended for training with distributional losses.

Normalization. BatchNorm is categorically incompatible with distributional losses: it enforces zero-mean unit-variance per activation within the batch, directly cancelling the inter-sample variance signal that ED and SD inject through the distributional gradient. Figure 10 demonstrates this on the ring task: SD with BatchNorm produces visibly fragmented, mode-collapsed outputs, while the same loss without normalization recovers the target distribution. The figure also shows three additional losses not in the main comparison: SDiv (recall-only) and NSD both recover the ring structure, while the NN-Cross term alone fails completely, confirming that it requires the profile comparison backbone of SD. LayerNorm (normalizing per sample, not per batch) is the safe alternative. No normalization at all is also viable when inputs are pre-normalized to unit variance.

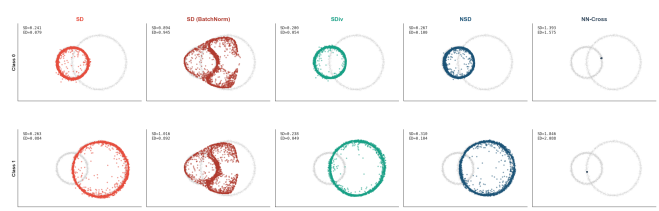

Fig. 10: Normalization and component analysis on the ring generation task (class 0 on top, class 1 on bottom). *Left pair:* SD succeeds without normalization but fails catastrophically with BatchNorm. BatchNorm normalises activations *across the batch*, cancelling the inter-sample variance signal that distributional losses depend on: the generated distribution fragments into disconnected clusters rather than forming a coherent ring. This demonstrates that BatchNorm and distributional losses are fundamentally incompatible. *Right panels:* component ablation. The recall term SDiv alone ($\mathcal{O}(nm)$ cost) recovers the ring, while the NN-Cross term alone (without profile comparison) collapses entirely, indicating that profile matching is the essential mechanism.

Gradient flow. Detaching the within-set signature profiles from autograd (treating S_{XX} and S_{YY} as zero-gradient constants) degrades performance significantly. The gradient flowing through S_{XX} provides the density-conditioning signal that prevents mode collapse.

The full PyTorch implementation, which supports full-matrix or batched pairwise distance computation, quantile alignment for unequally sized sets, and acceleration on GPUs and Apple Silicon, is provided in the accompanying repository.

Effect of batch composition. To illustrate the importance of the glocal protocol, Figures 12–13 show results from the same architecture and training setup, but with a global-only loss (applied to the full mixed-tissue batch without per-tissue terms).

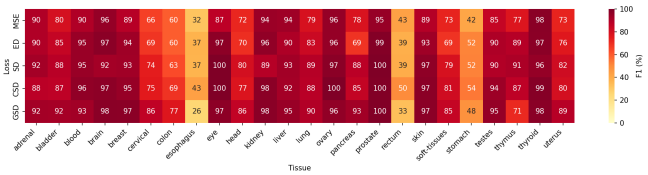

Fig. 11: Per-tissue F1 scores for the tissue-conditioned generation experiment (noise dim = 32, glocal protocol). Each cell shows the F1 score (%) for a given tissue–loss combination, with a classifier trained exclusively on synthetic data and evaluated on real test samples. The glocal protocol enables all losses to achieve non-trivial F1 across all 24 tissues. Distributional losses (ED, SD, CSD) consistently improve over MSE, particularly for tissues where MSE scores are lowest. No tissue is categorically failed by any loss, in contrast to the global-only training (Figure 12).

Under global-only training, MSE achieves 82.8% accuracy—the highest of all losses—because pointwise regression does not require distributional structure within each tissue. In contrast, the distributional losses collapse: SD achieves 49.1%, while CSD (10.0%) and GSD (14.5%) fall near chance level for 24 classes. With the glocal protocol (main text, Table 1), distributional losses overtake MSE and all exceed 83% accuracy. This demonstrates that batch composition is not merely a convenience but a prerequisite for effective use of distributional losses on multi-population data.

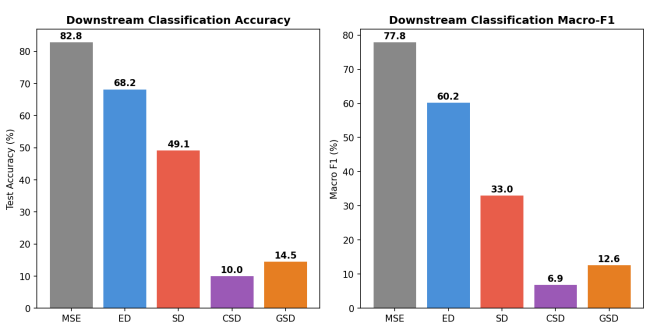

Fig. 12: Global-only training baseline: downstream tissue classification accuracy and macro-F1 using the same architecture and training setup as the main experiment, but without per-tissue local terms. Without the glocal protocol, MSE achieves the highest accuracy (82.8%) because pointwise regression implicitly pairs generated and real samples, requiring no within-tissue distributional structure. Distributional losses that operate on the full mixed-tissue batch collapse: SD (49.1%) retains moderate row-level discrimination, but CSD (10.0%) and GSD (14.5%) fall near chance. Compare with the glocal results (Table 1), where distributional losses overtake MSE and GSD leads with 89.9%.

Appendix C: Measure-Theoretic Formulation and the Negative Definite Conjecture

The discrete algorithm in Section 2.1 is the natural estimator of a continuous population-level quantity. Let μ and ν be probability

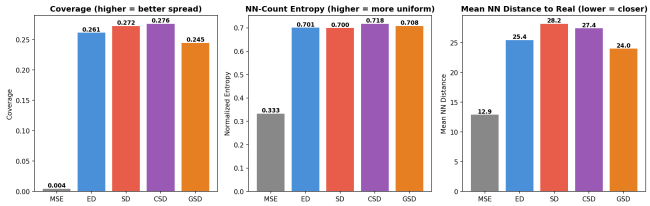

Fig. 13: Global-only training baseline: 1-NN distributional analysis. MSE achieves near-zero coverage and entropy (barycentric collapse to tissue means), while all distributional losses achieve comparable coverage and entropy. This confirms that distributional losses successfully diversify the generated distribution even under global-only training, but their poor *accuracy* (Figure 12) reveals that diversity without within-tissue structure does not translate to tissue-discriminative samples.

measures on $\mathbb{R}^d$. For any reference point z, the sorted signature of distances from z is the eCDF of radial distances; in the population limit, at radius r this becomes the CDF $F_{z\mu}(r)$, equal to the measure of the closed ball $B_r(z)$:

$$F_{S_z(\mu)}(r) = \mu\big(B_r(z)\big) \tag{14}$$

Because the W_1 distance between two one-dimensional distributions equals the L_1 integral of the absolute difference of their CDFs, the pointwise divergence at reference z is:

$$W_1\big(S_\mu(z),\, S_\nu(z)\big) = \int_0^\infty \big|\mu(B_r(z)) - \nu(B_r(z))\big|\, dr \tag{15}$$

Substituting into the symmetrized aggregate over all reference points, weighted by the barycentric mixture $\frac{\mu+\nu}{2}$, yields the continuous form of Squared Signature Distance:

$$\mathrm{SD}^2(\mu,\nu) = \int_{\mathbb{R}^d}\int_0^\infty \big|\mu(B_r(z)) - \nu(B_r(z))\big|\, dr\, d\!\left(\frac{\mu+\nu}{2}\right)(z) \tag{16}$$

This formulation makes precise the sense in which SD is a local projection of the optimal transport geometry: it measures how much the two distributions differ in their ball-mass assignments, integrated over all radii and reference locations.

We collect here the characterization of the open problem regarding the triangle inequality.

Schoenberg's Theorem. The precise route to proving the triangle inequality for $\mathrm{SD} = \sqrt{\mathrm{SD}^2}$ runs through a classical result of Schoenberg (1938). A symmetric kernel $K(\mu,\nu) \geq 0$ generates a valid metric via its square root if and only if K is of *conditionally negative definite* (CND) type: for every finite collection of measures $\mu_1, \ldots, \mu_n$ and real coefficients $c_1, \ldots, c_n$ with $\sum_i c_i = 0$,

$$\sum_{i,j} c_i c_j K(\mu_i, \mu_j) \leq 0 \tag{17}$$

Equivalently, CND kernels are those whose distance matrices, projected onto the zero-sum subspace, have non-positive eigenvalues. When K is CND, $\sqrt{K}$ embeds isometrically into a Hilbert space and satisfies the triangle inequality unconditionally.

SD^2 as a non-MMD divergence. Establishing the CND property is standard for energy distance, because $\mathcal{E}^2$ is a Maximum Mean Discrepancy (MMD): it decomposes as a linear

functional of the measures via a positive definite kernel. The continuous formulation in Equation 16 reveals why SD^2 is more difficult. The integration measure $d\big(\frac{\mu+\nu}{2}\big)(z)$ depends on both arguments simultaneously, making SD^2 non-linear with respect to its components. This non-linearity precludes the standard Hilbert space embedding argument that handles MMDs, and requires finding a non-linear Hilbert space embedding instead.

The global cancellation phenomenon. The standard strategy for proving that an integrated kernel is CND is to verify that the integrand is CND at every fixed evaluation parameter, then invoke closure under integration. Isolating the fixed-radius kernel at radius r:

$$\Delta_r(\mu,\nu) = \int_{\mathbb{R}^d} \big|\mu(B_r(z)) - \nu(B_r(z))\big|\, d\left(\frac{\mu+\nu}{2}\right)(z) \tag{18}$$

Extensive numerical eigenvalue testing reveals that $\Delta_r(\mu,\nu)$ is *not* conditionally negative definite for any fixed radius r: projecting the Gram matrix onto the zero-sum subspace yields strictly positive eigenvalues. Any proof strategy that proceeds pointwise in r is therefore structurally blocked.

Nevertheless, the distance matrix of $\mathrm{SD}^2 = \int_0^\infty \Delta_r\, dr$ is unconditionally CND across all tested configurations up to machine precision. The positive CND-violating eigenvalues of the fixed-radius kernels Δ_r cancel across the full radial integral. This *global cancellation phenomenon* suggests that any analytic proof of the CND property for SD^2 must exploit the full integral structure over $r \in [0,\infty)$ and cannot be reduced to a pointwise argument.

Identifying a functional-analytic mechanism that explains this global cancellation is the precise open problem. Its resolution would formally establish that $\mathrm{SD} = \sqrt{\mathrm{SD}^2}$ satisfies the triangle inequality unconditionally and embeds isometrically into Hilbert space.

Appendix D: Analytic Reduction to Energy Distance

Signature distance is a structural generalization of energy distance (Székely and Rizzo, 2013), defined as:

$$\mathcal{E}(X,Y) = 2\mathbb{E}\|X-Y\| - \mathbb{E}\|X-X'\| - \mathbb{E}\|Y-Y'\| \tag{19}$$

Consider the macro-separation regime where X and Y are completely disjoint, with minimum cross-distance exceeding their maximum internal diameters. The intra- and cross-signature arrays then do not overlap ($S_{XY}[k] \geq S_{XX}[k]$ for all k), so the absolute value in δ drops out: $\delta(S_{XX}, S_{XY}) = \mathrm{mean}(S_{XY}) - \mathrm{mean}(S_{XX})$. Substituting into Equation 4, the expression under the square root reduces to:

$$\mathrm{SD}^2(X,Y) = \mathbb{E}\|X-Y\| - \frac{1}{2}\mathbb{E}\|X-X'\| - \frac{1}{2}\mathbb{E}\|Y-Y'\| = \frac{1}{2}\mathcal{E}(X,Y) \tag{20}$$

Energy distance is therefore both a lower bound and an asymptotic limit of SD^2. The lower bound $\frac{1}{2}\mathcal{E}(X,Y) \leq \mathrm{SD}^2(X,Y)$ holds globally for any pair of distributions (Appendix A). Equality $\mathrm{SD}^2 = \frac{1}{2}\mathcal{E}$ holds in the macro-separation regime; for overlapping distributions, the non-linear sorting activates and SD^2 exceeds $\frac{1}{2}\mathcal{E}$, capturing local topological structure that the scalar mean-based reduction cannot resolve.

References

Argelaguet R, Velten B, Arnol D, et al. Multi-Omics Factor Analysis — a framework for unsupervised integration of multi-omics data sets. *Molecular Systems Biology*. 2018;14(6):e8124.

Beyer K, Goldstein J, Ramakrishnan R, Shaft U. When is "nearest neighbor" meaningful? *Database Theory — ICDT'99, Lecture Notes in Computer Science*. 1999;1540:217–235.

Carlsson G. Topology and data. *Bulletin of the American Mathematical Society*. 2009;46(2):255–308.

Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*. 2002;16:321–357.

Cui H, Wang C, Maan H, et al. scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nature Methods*. 2024;21(8):1470–1480.

Cuturi M. Sinkhorn distances: lightspeed computation of optimal transport. *Advances in Neural Information Processing Systems*. 2013;26:2292–2300.

Dixit A, Parnas O, Li B, et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell*. 2016;167(7):1853–1866.

Edelsbrunner H, Harer JL. Computational Topology: An Introduction. American Mathematical Society; 2010.

Efron B, Tibshirani RJ. An Introduction to the Bootstrap. Chapman & Hall/CRC; 1993.

Ernst MD. Permutation methods: a basis for exact inference. *Statistical Science*. 2004;19(4):676–685.

Goodfellow I, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets. *Advances in Neural Information Processing Systems*. 2014;27:2672–2680.

Gretton A, Borgwardt KM, Rasch MJ, Schölkopf B, Smola AJ. A kernel two-sample test. *Journal of Machine Learning Research*. 2012;13:723–773.

Heusel M, Ramsauer H, Unterthiner T, Nessler B, Hochreiter S. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. *Advances in Neural Information Processing Systems*. 2017;30.

Kingma DP, Welling M. Auto-encoding variational Bayes. *International Conference on Learning Representations*. 2014.

Kuhn HW. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*. 1955;2(1–2):83–97.

Kynkäänniemi T, Karras T, Laine S, Lehtinen J, Aila T. Improved precision and recall metric for assessing generative models. *Advances in Neural Information Processing Systems*. 2019;32.

Lloyd S. Least squares quantization in PCM. *IEEE Transactions on Information Theory*. 1982;28(2):129–137.

Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. *Nature Methods*. 2018;15(12):1053–1058.

Luecken MD, Theis FJ. Current best practices in single-cell RNA-seq analysis: a tutorial. *Molecular Systems Biology*. 2019;15(6):e8746.

Marouf M, Machart P, Bansal V, et al. Realistic in silico generation and augmentation of single-cell RNA-seq data using generative adversarial networks. *Nature Communications*. 2020;11(1):166.

Peyré G, Cuturi M. Computational optimal transport. *Foundations and Trends in Machine Learning*. 2019;11(5–6):355–607.

Phipson B, Smyth GK. Permutation P-values should never be zero: calculating exact P-values when permutations are randomly

drawn. *Statistical Applications in Genetics and Molecular Biology*. 2010;9(1):Article 39.

Rezende DJ, Mohamed S. Variational inference with normalizing flows. *International Conference on Machine Learning*. 2015:1530–1538.

Schoenberg IJ. Metric spaces and completely monotone functions. *Annals of Mathematics*. 1938;39(4):811–841.

Subramanian A, Narayan R, Corsello SM, et al. A next generation connectivity map: L1000 platform and the first 1,000,000 profiles. *Cell*. 2017;171(6):1437–1452.

Székely GJ, Rizzo ML. Energy statistics: A class of statistics based on distances. *Journal of Statistical Planning and Inference*. 2013;143(8):1249–1272.

The Cancer Genome Atlas Research Network. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*. 2008;455(7216):1061–1068.

Theodoris CV, Xiao L, Chopra A, et al. Transfer learning enables predictions in network biology. *Nature*. 2023;618(7965):616–624.

Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*. 2019;9:5233.

Villani C. Optimal Transport: Old and New. Springer; 2009.

Welling M, Teh YW. Bayesian learning via stochastic gradient Langevin dynamics. *International Conference on Machine Learning*. 2011:681–688.