Riemannian generative decoder

Andreas Bjerregaard [1] **Søren Hauberg** [2] **Anders Krogh** [1]

Abstract

Riemannian representation learning typically relies on approximating densities on chosen manifolds. This involves optimizing difficult objectives, potentially harming models. To completely circumvent this issue, we introduce the Riemannian generative decoder which finds manifold-valued maximum likelihood latents with a Riemannian optimizer while training a decoder network. By discarding the encoder, we vastly simplify the manifold constraint compared to current approaches which often only handle few specific manifolds. We validate our approach on three case studies — a synthetic branching diffusion process, human migrations inferred from mitochondrial DNA, and cells undergoing a cell division cycle — each showing that learned representations respect the prescribed geometry and capture intrinsic non-Euclidean structure. Our method requires only a decoder, is compatible with existing architectures, and yields interpretable latent spaces aligned with data geometry.

Home yhsure.github.io/riemannian-generative-decoder
Code github.com/yhsure/riemannian-generative-decoder
Data hf.co/datasets/yhsure/riemannian-generative-decoder

1. Introduction

Real-world data often lie on non-Euclidean manifolds — e.g., evolutionary trees, social-network graphs, or periodic signals — yet most latent-variable models assume $\mathbb{R}^d$. Traditional Euclidean methods often fail to provide visualizations rooted in the geometry we know underlies the data. Riemannian manifolds provide a natural framework for modeling geometrical structures. Despite the flexibility of variational autoencoders, enforcing manifold priors (e.g. von Mises–Fisher on spheres or Riemannian normals in hyperbolic spaces) requires complex densities and Monte Carlo estimates of normalizing constants, limiting scalability to general manifolds.

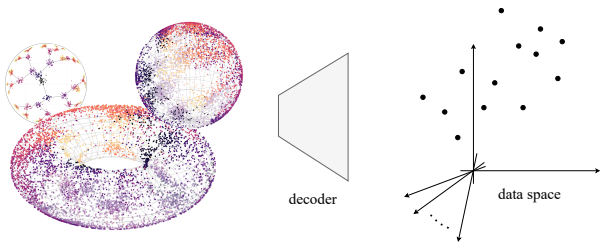


Figure 1: Methodological overview; a decoder reconstructs data from Riemannian manifolds where representations are optimized as model parameters via maximum likelihood.

We propose the *Riemannian generative decoder*: we discard the encoder and directly learn manifold-valued latents with a Riemannian optimizer while training a decoder network. This encoderless scheme removes the need for closed-form densities or approximate encodings on the manifold, and handles any Riemannian manifold. By using a *geometric* flavor of regularization through input noise, our model is further encouraged to penalize large output gradients relative to the local curvature. We analyze this form of regularization and see its importance in preserving geometric structure during dimensionality reduction.

2. Background

Learned representations often reveal the driving patterns of the data-generating phenomenon. Much of computational biology — and especially data-driven fields like transcriptomics — greatly rely on dimensionality reduction techniques to understand the underlying factors of their experiments (Becht et al., 2019). Unfortunately, a lack of statistical identifiability implies that such representations need not be unique (Locatello et al., 2019). Therefore, it is common practice to inject various inductive biases that reflect prior beliefs or hypotheses about the analyzed problem.

2.1. Latent variable models

Autoencoders (AEs) learn a deterministic mapping $x \mapsto z \mapsto \hat{x}$ by minimizing a reconstruction loss

[1]University of Copenhagen, Denmark [2]Technical University of Denmark, Denmark. Correspondence to: Andreas Bjerregaard <anje@di.ku.dk>, Anders Krogh <akrogh@di.ku.dk>.

Proceedings of the Workshop on Generative AI for Biology at the 42nd International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

$$\min_{\theta,\phi} \sum_{i=1}^{N} L\left(x_i, f_\theta(g_\phi(x_i))\right)$$

where L could measure squared error, g_ϕ is the encoder and f_θ the decoder. Because f_θ is typically smooth, nearby latent codes produce similar reconstructions. This imposes a *smoothness bias* on the representation: distances in latent space are tied to distances in data space.

The variational autoencoder (VAE) by Kingma et al. (2013) extends this by introducing a prior $p(z)$ and a stochastic encoder $q_\phi(z\,|\,x)$ as a variational distribution. The intractable marginal likelihood

$$p(x|\theta) = \int p(x|z,\theta)p(z)dz$$

is then lower bounded via

$$\log p(x \mid \theta) \;\geq\; \underbrace{\mathbb{E}_{q_\phi(z|x)}\big[\log p_\theta(x \mid z)\big] \;-\; D_{\mathrm{KL}}\big(q_\phi(z \mid x) \,\|\, p(z)\big)}_{\mathrm{ELBO}(x;\theta,\phi)} \tag{1}$$

The decoder is trained by maximizing the ELBO to reconstruct x from samples of $z \sim q_\phi(z \mid x)$. The KL term encourages the encoding distribution to match the prior, typically $\mathcal{N}(0, I)$, while the stochasticity of q_ϕ forces the decoder to be robust to perturbations in z. Together, these constraints strengthen the smoothness bias across the encoder distribution.

An alternative to the VAE is the Deep Generative Decoder (Schuster and Krogh, 2023), avoiding an encoder entirely. Each latent z_i is treated as a free parameter, and the model maximizes

$$\sum_{i=1}^{N} \log p_\theta(x_i \mid z_i) + \log p(z_i \mid \phi)$$

via *maximum likelihood* or *maximum a posteriori* estimation. A parameterized distribution $p(z_i \mid \phi)$ on latent space, such as a Gaussian mixture model, can introduce inductive bias. The decoder smoothness imposes again a continuity constraint on $z \mapsto x$, as reconstructions must interpolate well across learned codes. Unlike the VAE, no amortized inference is used, but the same decoder regularization implicitly shapes the latent geometry.

In all three frameworks — AE, VAE, and DGD — the smoothness of the decoder function acts as a regularizer on latent codes. Since nearby z produce similar outputs, the learned representations inherit geometric continuity. The VAE further strengthens this bias through stochastic encodings and KL regularization. The DGD enforces it by directly optimizing per-sample codes under a smooth decoder. These smoothness priors play a central role in learning meaningful low-dimensional structure.

2.2. Geometric biases

Most learned representations are assumed to be Euclidean, belonging to $\mathbb{R}^d$. This implies a simple, unbounded topological structure for the representations. This is a flexible and not very informative inductive bias. We briefly survey parts of the literature and generally find that existing approaches involve layers of complexity that potentially limit their performance.

Spherical representation spaces encode compactness and periodicity. Davidson et al. (2018) and Xu and Durrett (2018) define latents on $\mathbb{S}^{d-1}$ via a von Mises–Fisher prior:

$$p(z \mid \mu, \kappa) = C_d(\kappa)\, \exp\big(\kappa\, \mu^\top z\big) \tag{2}$$
$$\text{with}\;\; C_d(\kappa) = \frac{\kappa^{d/2-1}}{(2\pi)^{d/2}\, I_{d/2-1}(\kappa)}$$

where $\mu \in \mathbb{S}^{d-1}$ and $\kappa > 0$. Sampling uses rejection or implicit reparameterization and KL terms involve Bessel functions, complicating Equation 1 while adding computational overhead and bias.

Hyperbolic representation spaces effectively capture hierarchical data structures (Krioukov et al., 2010). A popular choice (used for, e.g., the $\mathcal{P}$-VAE (Mathieu et al., 2019)) is the Poincaré ball $\mathbb{B}^d$, with metric $g_z = \lambda(z)^2 I$, where $\lambda(z) = 2/(1 - \|z\|^2)$, and distance

$$d_{\mathbb{B}}(u, v) = \operatorname{arcosh}\Big(1 + 2\frac{\|u - v\|^2}{(1 - \|u\|^2)(1 - \|v\|^2)}\Big).$$

One typically uses the Riemannian normal prior

$$p(z) \propto \exp\big(-d_{\mathbb{B}}(z, \mu)^2/(2\sigma^2)\big). \tag{3}$$

The ELBO then requires approximating both the intractable normalizing constant of the prior and volume corrections, typically via Monte Carlo or series-expansion methods. Alternative hyperbolic embeddings like the Lorentz (Nickel and Kiela, 2018) or stereographic projections (Skopek et al., 2019) improve computational stability and flexibility but face analogous challenges.

General geometries can represent different inductive biases (Kalatzis et al., 2020; Connor et al., 2021; Falorsi et al., 2018; Grattarola et al., 2019). Current literature is based on encoders whose densities generally lack closed-form formulas on arbitrary manifolds $\mathcal{M}$, relying on approximations like Monte Carlo importance sampling, truncated wrapped normals

$$q(z \mid \mu, \Sigma) \approx \sum_{k \in \mathbb{Z}^d} \frac{\exp\big(-\frac{1}{2}\| \operatorname{Log}_\mu(z) + 2\pi k\|^2_{\Sigma^{-1}}\big)}{(2\pi)^{d/2}\, |\Sigma|^{1/2}} \tag{4}$$

or random-walk reparameterization encoders such as ΔVAE (Rey et al., 2019), simulating Brownian motion using the manifold exponential map: $z = \operatorname{Exp}_\mu(\sum_i \xi_i)$, $\xi_i \sim \mathcal{N}(0, \frac{t}{\text{steps}} I)$.

3. Methodology

Much of the difficulty in probabilistically learning representations over non-trivial geometries is that densities are notably more difficult to handle than, e.g., a Gaussian distribution. Inspired by Schuster and Krogh (2023) we propose not to variationally infer the representations, but instead perform standard maximum likelihood estimation (MLE). Combining Riemannian optimization with a form of geometry-aware regularization, we build a simple yet effective representation learning scheme that works across different geometric inductive biases.

3.1. Model formulation

Let $x \in \mathcal{X}$ denote observed data samples, and let $z \in \mathcal{M}$ represent latent variables constrained to a Riemannian manifold $\mathcal{M}$. We define a decoder function $f_\theta : \mathcal{M} \to \mathcal{X}$, parameterized by θ, which maps latent representations to the data space. With a reconstruction metric, we perform MLE of the decoder parameters θ and the latent representations $Z = \{z_1, z_2, \ldots, z_N\}$ for a dataset $X = \{x_1, x_2, \ldots, x_N\}$. Assuming i.i.d. samples and applying the logarithm then gives us the training objective:

$$\begin{aligned} &\arg\max_{Z,\theta} p(X \mid Z, \theta)\, p(Z) \\ &= \arg\max_{Z,\theta} \sum_{i=1}^{N} \big(\log p(x_i \mid z_i, \theta) + \log p(z_i)\big) \end{aligned} \tag{5}$$

For compact manifolds, we assume a uniform distribution $\log p(z_i) = -\log \mathrm{Vol}(\mathcal{M})$ and $\log p(z_i)$ need not be included in the loss. At generation time, one samples $z \sim \mathrm{Uniform}(\mathcal{M})$ to compute $x = f_\theta(z)$. For non-compact manifolds, choices like the wrapped normal can function as $p(z)$.

To enforce the manifold constraint on Z, we use Riemannian Adam (Bécigneul and Ganea, 2018) on the latent representations. Such optimizers work by following the Riemannian gradient projected onto the tangent space, and mapping the updated point back onto the manifold. Relying on *geoopt* (Kochurov et al., 2020) for defining tensors and optimizing on manifolds, the implementation becomes exceedingly simple (see Appendix A).

For manifolds whose metric tensor varies with position, we introduce a *geometry-aware regularization*. During training, each latent z is perturbed with Gaussian noise whose covariance is the inverse Riemannian metric at z. This adapts the noise to local curvature: on homogeneous manifolds such as the hypersphere (where curvature is constant and metric variation merely reflects coordinate scaling) the procedure recovers nearly isotropic noise, whereas on spaces with non-uniform curvature the noise shape is greatly adjusted. We outline a derivation inspired by Bishop (1995) and An (1996) to analyze this noise:

Let $\epsilon \sim \mathcal{N}\big(0, \Sigma\, G^{-1}(z)\big)$ and define the squared-error loss $L(z) = \|f(z,\theta) - y\|^2$ for some target y. We inject noise via the exponential map, which we approximate by the identity to $O(\|\epsilon\|^2)$:

$$z' = \mathrm{Exp}_z(\epsilon) = z + \epsilon + O(\|\epsilon\|^2).$$

Ignoring higher order terms, a second-order Taylor expansion around z gives

$$L(z') \approx L(z) + \nabla_z L(z)^\top \epsilon + \tfrac{1}{2}\, \epsilon^\top H_z L(z)\, \epsilon,$$

where $H_z L(z) = \nabla_z^2 L(z)$. Taking expectation over ϵ and using $\mathbb{E}[\epsilon] = 0$, $\mathbb{E}[\epsilon\epsilon^\top] = \Sigma\, G^{-1}$, we obtain

$$\mathbb{E}[L(z')] = L(z) + \tfrac{1}{2}\ \mathrm{Tr}(H_z L(z)\, \Sigma\, G^{-1}).$$

For squared error, we have

$$H_z L(z) = 2J(z)^\top J(z) + \sum_k (f_k(z) - y_k)\, H_z f_k(z),$$

with $J(z) = \partial_z f(z,\theta)$. For non-biased estimates $f_k(z)$ the residual in the second term is negligible on average, so

$$\mathbb{E}[L(z')] \approx L(z) + \mathrm{Tr}(J^\top \Sigma\, G^{-1} J).$$

Thus corrupting representations with noise scaled with G^{-1} induces a regularizer $\mathrm{Tr}(J^\top \Sigma\, G^{-1} J)$. In our study, we consider the special case of isotropic Gaussian noise, i.e., $\Sigma = \sigma^2 I$. In this setting, the regularization term simplifies to

$$\mathrm{Tr}\big(J^\top \Sigma\, G^{-1} J\big) = \sigma^2\ \mathrm{Tr}\big(J^\top G^{-1} J\big). \tag{6}$$

3.2. Datasets

Cell cycle stages. Measuring gene expression of individual fibroblasts with single-cell RNA sequencing captures a continuous, asynchronous progression through the cell division cycle. Transcriptomic changes occur through these phases, yielding cyclic patterns in gene expression. As the data is not coupled in nature (we cannot identify and keep track of individual cells), unsupervised learning is suitable for picking up patterns about the underlying distribution of cells. We apply our Riemannian generative decoder to the human fibroblast scRNA-seq dataset (5367 cells × 10789 genes) introduced in DeepCycle (Riba et al., 2022) and archived on Zenodo (Riba, 2021). Data were already preprocessed by scaling each cell to equal library size, log-transforming gene counts, and smoothing and filtering using a standard single-cell pipeline. Before modeling, we subsampled to 189 genes annotated with the cell cycle gene ontology term (GO:0007049) retrieved via QuickGO (Binns et al., 2009) in accordance with other cell cycle studies.

Branching diffusion process. The synthetic dataset from Mathieu et al. (2019)[1] simulates data with tree-like structure using a hierarchical branching diffusion process. Starting from a root node at the origin in $\mathbb{R}^d$, we generate a tree of total depth D, where each node at depth ℓ produces C child nodes at depth $\ell + 1$. Child nodes are obtained by adding Gaussian noise to their parent node, with variance *optionally* decreasing with depth:

$$x_{\text{child}} = x_{\text{parent}} + \epsilon, \quad \epsilon \sim \mathcal{N}\left(0, \frac{\sigma^2}{p^\ell} I\right),$$

where σ is the base noise level, $p \geq 1$ is a scaling parameter, and I is the identity matrix. For every node, we also generate S noisy sibling observations to simulate measurement noise:

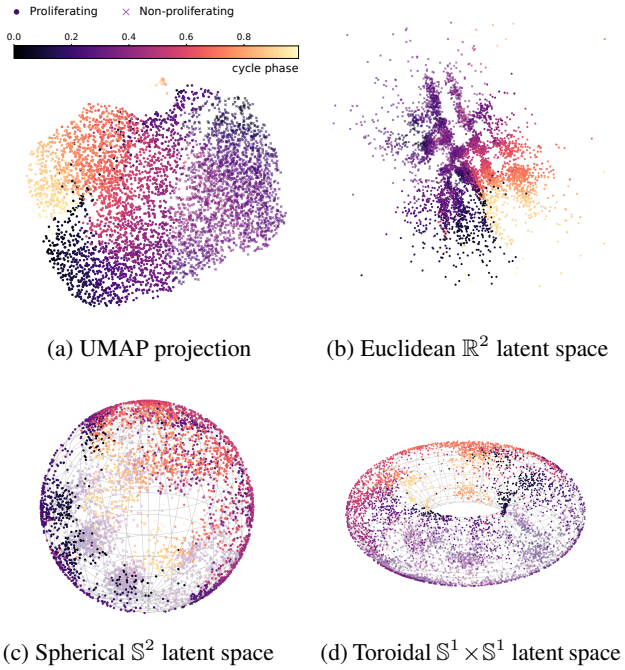
$$x_{\text{obs}} = x_{\text{node}} + \epsilon', \quad \epsilon' \sim \mathcal{N}\left(0, \frac{\sigma^2}{f p^\ell} I\right),$$

where $f > 1$ controls the observation noise level. The dataset comprises all noisy observations x_{obs}, inherently containing the hierarchical relationships from the branching process. Prior to training, data is standardized to have zero mean and unit variance. We set $d = 50, D = 7, C = 2, \sigma = 1, p = 1$ and generate $S = 50$ noisy sibling observations per node with an observation-noise factor $f = 8$; this generates 6 350 noisy observations.

Human mitochondrial DNA. Human mitochondrial DNA (hmtDNA) is a small, maternally inherited genome found in cells' mitochondria. Its relatively compact size and stable inheritance make it a fundamental genetic marker in studies of human evolution and population structure. A relatively rapid mutation rate has led the genomes to distinct genetic variants, named *haplogroups*, which reflect evolutionary branching events.

We retrieved 67 305 complete or near-complete sequences (15 400 – 16 700 bp) from GenBank via a query from MITOMAP (MITOMAP, 2023). Sequences were annotated with haplogroup labels using *Haplogrep3* (Schönherr et al., 2023), leveraging phylogenetic trees from PhyloTree Build 17 (Van Oven, 2015). A sequence was kept if the reported quality was higher than 0.9. In addition to haplogroup classification, *Haplogrep3* identified mutations with respect to a root sequence; here, separate datasets were made using either the rCRS (revised Cambridge reference sequence) or RSRS (reconstructed Sapiens reference sequence). Mutations were then encoded in a one-hot scheme, removing mutations with ≤ 0.05 frequency, resulting in datasets with shapes 61665×6298 (rCRS) and 57385×5366 (RSRS). Appendix D displays further characteristics.

[1]Available under MIT license



(a) UMAP projection (b) Euclidean $\mathbb{R}^2$ latent space

(c) Spherical $\mathbb{S}^2$ latent space (d) Toroidal $\mathbb{S}^1 \times \mathbb{S}^1$ latent space

Figure 2: **Cell cycle phases using either (a) UMAP or (b–d) different Riemannian manifolds.** Samples are concatenated across train/validation/test sets. The phase is inferred by DeepCycle as a continuous variable $\phi \in [0, 1)$ which wraps around such that $\phi = 0$ and $\lim_{\phi \to 1} \phi$ denote the same point in the cycle.

4. Results and discussion

In the following, we treat each dataset to evaluate and discuss applications of unsupervised learning on meaningful geometries.

4.1. Cell cycle stages

Figure 2 shows the resulting representations from learning generative models with different manifolds on data with a underlying cyclical property. We see that neural networks are difficult to control. While we may have an idea of an explainable global optimum — e.g., a neatly arranged circle following the cell cycle stages — optimization of neural networks cannot easily be controlled to follow such an idea. Given a model expressive enough, representations lying in a circle could as well be unrolled or have distinct arcs interchanged without any loss in task accuracy.

To compare model fidelity and how well manifold distances correspond to distance in cell cycle, Table 1 lists reconstruction fidelities and correlations between phase distances and manifold geodesics. Here we compared to $\mathcal{S}$-VAE (Davidson et al., 2018) and ΔVAE (Rey et al., 2019); see Appendix C for further details. Euclidean $\mathbb{R}^3$ with 3 degrees of freedom achieves better reconstructions than the other

Table 1: **Cell cycle: Correlation and reconstruction metrics across five random initializations** (formatted as mean ± std). Pearson/Spearman correlate phase distances to latent distances while MAE/MSE measure reconstruction by L1/L2-norm. Our models in gray. Comparisons with $\mathcal{S}$-VAE (Davidson et al., 2018) and ΔVAE (Rey et al., 2019).

	Train				Test			
	Pearson	Spearman	MAE	MSE	Pearson	Spearman	MAE	MSE
Euclidean $\mathbb{R}^2$	0.47±0.03	0.50±0.03	0.31±0.00	0.17±0.00	0.52±0.03	0.53±0.03	0.31±0.00	0.18±0.00
Euclidean $\mathbb{R}^3$	0.50±0.05	0.54±0.04	**0.30**±0.00	**0.16**±0.00	0.55±0.04	0.57±0.03	**0.31**±0.00	**0.17**±0.00
Sphere $\mathbb{S}^2$	**0.58**±0.03	**0.59**±0.03	0.31±0.00	0.17±0.00	**0.60**±0.03	**0.60**±0.03	0.32±0.00	0.18±0.00
Torus $\mathbb{S}^1 \times \mathbb{S}^1$	0.50±0.07	0.51±0.07	0.31±0.00	0.17±0.00	0.52±0.07	0.53±0.08	0.32±0.00	0.18±0.00
$\mathcal{S}$-VAE sphere	0.50±0.02	0.53±0.03	0.32±0.00	0.19±0.00	0.53±0.02	0.55±0.02	0.32±0.00	0.19±0.00
ΔVAE sphere	0.52±0.01	0.55±0.01	0.31±0.00	0.17±0.00	0.57±0.02	0.59±0.02	0.32±0.00	0.18±0.00
ΔVAE torus	0.43±0.07	0.45±0.07	0.31±0.00	0.17±0.00	0.48±0.06	0.50±0.07	0.32±0.00	0.18±0.00

Table 2: **Branching diffusion: Correlation and reconstruction metrics across five random initializations** (formatted as mean ± std). Pearson and Spearman correlate all pairs of distances in the tree structure with latent geodesic distances. Our models in gray. Comparison with $\mathcal{P}$-VAE (Mathieu et al., 2019).

	Train				Test			
	Pearson	Spearman	MAE	MSE	Pearson	Spearman	MAE	MSE
Euclidean $\mathbb{R}^2$ ($\sigma = 0.0$)	*0.53*±0.01	*0.49*±0.01	***0.14***±0.00	***0.03***±0.00	*0.52*±0.02	*0.49*±0.02	*0.18*±0.01	*0.06*±0.01
Sphere $\mathbb{S}^2$ ($\sigma = 0.0$)	*0.56*±0.02	*0.53*±0.03	***0.14***±0.00	***0.03***±0.00	*0.55*±0.02	*0.52*±0.02	***0.17***±0.00	***0.05***±0.00
Lorentz $\mathbb{H}^2$ ($\sigma = 0.1$)	0.52±0.02	0.48±0.02	0.15±0.00	0.04±0.00	0.48±0.02	0.45±0.03	0.19±0.02	0.08±0.02
Lorentz $\mathbb{H}^2$ ($\sigma = 0.5$)	0.78±0.02	0.74±0.03	0.32±0.01	0.18±0.02	0.69±0.03	0.69±0.03	0.28±0.02	0.14±0.02
Lorentz $\mathbb{H}^2$ ($\sigma = 1.0$)	**0.81**±0.02	**0.77**±0.02	0.49±0.01	0.39±0.01	**0.80**±0.02	**0.76**±0.02	0.36±0.01	0.21±0.01
Lorentz $\mathbb{H}^2$ ($\sigma = 2.0$)	0.77±0.04	0.74±0.05	0.68±0.01	0.74±0.01	0.79±0.09	0.73±0.11	0.52±0.02	0.45±0.03
$\mathcal{P}$-VAE $\mathbb{B}^2$ ($c = 1.2$)	0.68±0.03	0.54±0.07	0.42±0.02	0.30±0.02	0.68±0.04	0.54±0.09	0.42±0.02	0.31±0.02

manifolds as expected. Results from using tori (see Table 1) show a larger degree of variation than the sphere or a flat Euclidean space, suggesting this manifold is more sensitive to training dynamics. Learning on circles $\mathbb{S}^1$ embedded in 2D is likely not very flexible, causing these instabilities.

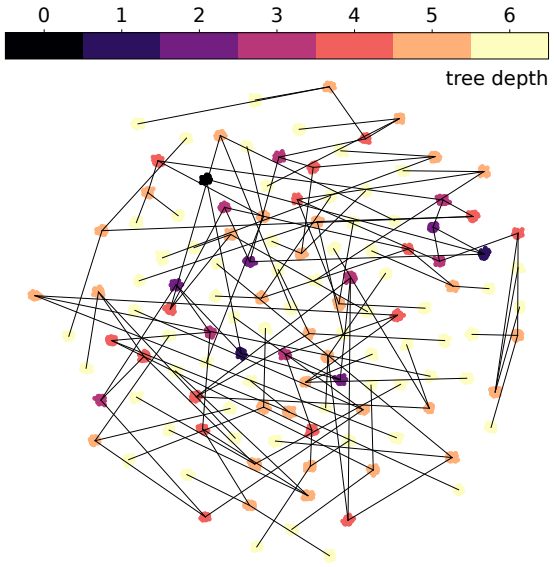
While our method significantly outperforms other models on the training data, the test results are generally similar across methods from different studies.[2] As implementation and optimization details differ slightly across studies — and considering some methods may be more robust to optimization hyperparameters than others — this area warrants further progress in benchmarking. Our work has focused on applications in bioinformatics, which has relevant real-world settings where distinct geometries occur and interpretability can lead to novel insights. For strict benchmarking, more synthetic datasets with controllable 'ground truth' geometries would help alleviate issues in cross-study comparisons.

[2]Used as a dimensionality reduction technique, generalization performance is generally of little significance.

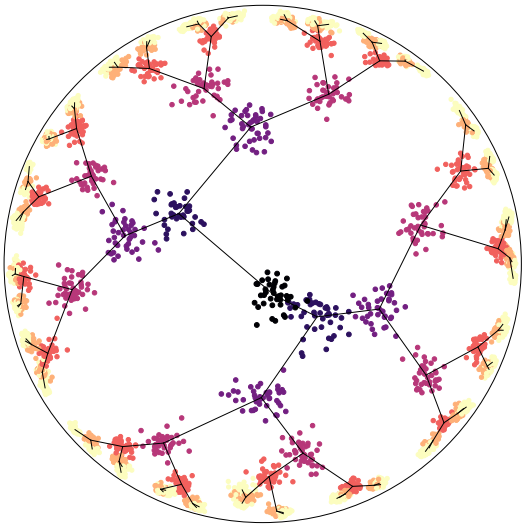
4.2. Branching diffusion process

Hyperbolic spaces can efficiently be used as a tool to uncover hierarchical processes. Notably, the UMAP projection (Figure 3a) fails to reveal any underlying tree topology, although yielding perfectly separated clusters for each tree node's noisy observations. In contrast, the regularized Poincaré disk embedding exhibits a clear structure mirroring the underlying hierarchical process (Figure 3b).

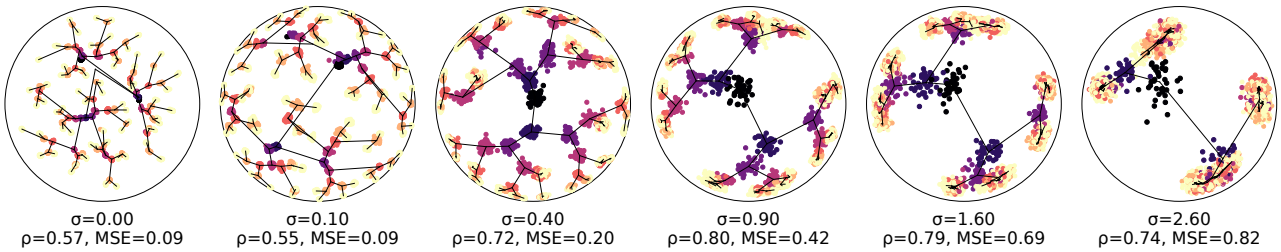
To study the effect of geometric regularization, Figure 3c shows models trained by fixing curvature $c = 5.0$ and varying noise from $\sigma = 0$ to $\sigma = 2.6$. Correlations increase sharply up to $\sigma \approx 0.9$, beyond which the noise completely overwhelms the decoder's capacity to preserve pairwise distances, starting a decline in the metrics. This indicates an optimal noise regime where geometric regularization helps structure latents according to the manifold's curvature,



(a) UMAP projection



(b) Poincaré projection ($\sigma = 0.5$)



(c) Progressively increasing regularization noise of the training process (6 distinct models)

Figure 3: **Visualizations of the branching diffusion process.** Trees consist of 7 levels with color lightness denoting depth. (a) UMAP projection; (b) Poincaré disk projection of Lorentz latents using geometric regularization ($c = 5.0, \sigma = 0.5$); (c) Ablation study showing the influence of the noise scale σ, listing Pearson correlation ρ and mean squared error on the training set.

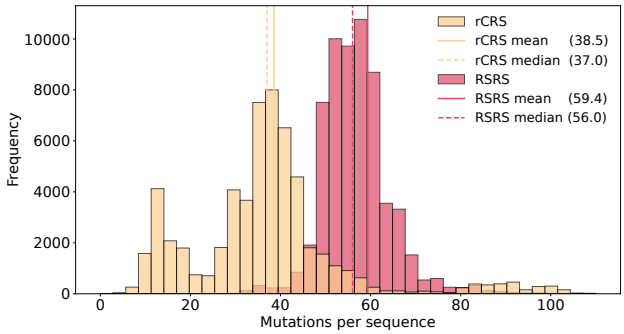


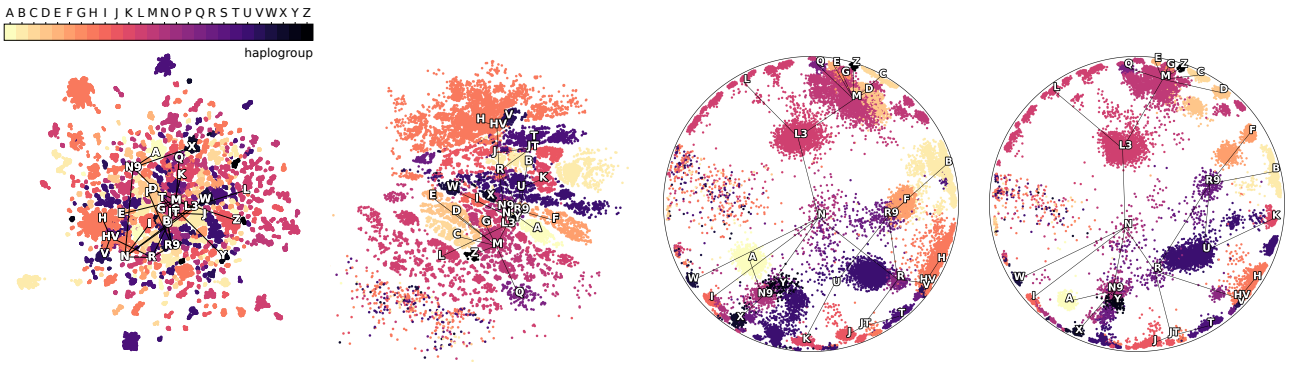
Figure 4: **Distributions of mutation counts per sequence**, comparing datasets encoded against the revised Cambridge reference sequence (rCRS) or the reconstructed Sapiens reference sequence (RSRS).

without excessively corrupting representations. Appendix F analyzes and shows how curvature and noise level relate to each other.

The results suggest a clear tradeoff between reconstruction and geometry. Future work could investigate adaptive schedules for σ, as well as tuning the ambient curvature constant c, to further optimize the balance between preservation of local structure and reconstruction fidelity.

4.3. Tracing human migrations from hmtDNA

By examining differences in hmtDNA sequences, it is possible to infer patterns of migration, lineage, and ancestry. In Figure 5, we show simplified lineages based on Lott et al. (2013) and PhyloTree (Van Oven, 2015). The figure shows how UMAP fails in uncovering the hierarchical nature (panel *a*), the Euclidean latents improve slightly (panel



(a) RSRS UMAP projection (b) RSRS Euclidean $\mathbb{R}^2$ (c) RSRS Poincaré ($\sigma = 0.5$) (d) rCRS Poincaré ($\sigma = 0.5$)

Figure 5: **Visualizations of hmtDNA haplogroups using either (a) UMAP, (b) Euclidean latent space, or (c–d) Poincaré projection of Lorentz latents** ($c = 5.0, \sigma = 0.5$)**.** Edges represent simplified lineage (Lott et al., 2013), nodes indicate median haplogroup positions.

Table 3: **Correlation and reconstruction train metrics across three runs for the hmtDNA dataset** (mean ± std). Mean F1 scores assess reconstruction; Pearson and Spearman correlate manifold vs genetic distance (5000 random points). RSRS/rCRS denote distinct reference sequences.

Ref.	Model	Pearson	Spearman	F1
RSRS	$\mathbb{H}^2$ ($\sigma = 0.1$)	0.15±0.01	0.12±0.04	0.93±0.00
RSRS	$\mathbb{H}^2$ ($\sigma = 0.5$)	0.28±0.01	**0.49**±0.02	0.86±0.00
RSRS	$\mathbb{R}^2$ ($\sigma = 0.0$)	**0.35**±0.00	0.29±0.01	**0.94**±0.00
rCRS	$\mathbb{H}^2$ ($\sigma = 0.1$)	0.18±0.02	0.17±0.05	0.88±0.00
rCRS	$\mathbb{H}^2$ ($\sigma = 0.5$)	0.28±0.01	**0.50**±0.04	0.79±0.01
rCRS	$\mathbb{R}^2$ ($\sigma = 0.0$)	**0.41**±0.03	0.42±0.10	**0.90**±0.00

b) while the hierarchy is much clearer for hyperbolic latents (panels *c–d*). Although the choice of reference sequence has a big impact on the actual mutational encoding of sequences (see Figure 4), models based on either reference converge at strikingly similar representations when using the same seed (comparing Figures 5c and d). The geographical locations of haplogroups strongly correspond to the locations of representations when comparing with maps from, e.g., Lott et al. (2013) — this may be quantified in future work.

For the geometry-aware Lorentz manifolds (middle rows of Table 3 for both reference sequences), Spearman correlations grow much larger than Pearson ones. This unveils a correlation which is *non-linear*. Intuitively, the manifold succeeds in capturing the hierarchical branching structure of the haplogroup tree, but the absolute path lengths are rescaled by the curvature. In this setting, we are mainly interesting in monotonicity as measured by Pearson. Note that the correlations of Table 3 are based on the distance in the one-hot data space; if one were to use a similar metric as from Table 2 — based on the distortion of the tree — then Figure 5 strongly suggests that the hyperbolic distances better relate to the tree structure than Euclidean ones.

Conclusions and future directions

By eliminating the encoder and utilizing Riemannian optimization and a geometry-aware regularization, we developed a new approach for Riemannian representation learning. Empirical validation demonstrated that our decoder successfully captures intrinsic geometric structures across diverse datasets, substantially improving correlations between latent distances and ground truth geometry. While we studied simple, low-dimensional manifolds in an interpretability setting, scaling to higher dimensions and more complex manifold combinations could lead to new results. Other research directions include adaptive geometric regularization strategies, extensions to manifold-valued network weights, and exploring latent manifold structures within pretrained neural networks (e.g., for generative diffusion processes or language models).

The decoder-only framework stores each representation explicitly, yielding memory that grows linearly with dataset size. Although Riemannian Adam enables efficient manifold-constrained updates, the per-sample parameterization may be prohibitive for datasets of millions of points. Hybrid schemes — such as amortized inference or low-rank factorization — could mitigate this. Lastly, our curated hmtDNA dataset invites for further empirical studies, including analyses of geographic distances, migration patterns, or distortion-based metrics of consensus trees. We make the data easily available on hf.co/datasets/yhsure/riemannian-generative-decoder.

Acknowledgements The authors thank Viktoria Schuster, Iñigo Prada-Luengo, Yan Li, Adrián Sousa-Poza and

Valentina Sora for helpful discussions. This work was funded by the Novo Nordisk Foundation (NNF) through the Center for Basic Machine Learning in Life Science (NNF grant NNF20OC0062606) and the Pioneer Centre for AI (DNRF grant number P1). A.K. was further supported by NNF grants NNF20OC0059939 and NNF20OC0063268.

References

G. An. The effects of adding noise during backpropagation training on a generalization performance. *Neural computation*, 8(3):643–674, 1996.

E. Becht, L. McInnes, J. Healy, C.-A. Dutertre, I. W. Kwok, L. G. Ng, F. Ginhoux, and E. W. Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38–44, 2019.

G. Bécigneul and O.-E. Ganea. Riemannian adaptive optimization methods. *arXiv preprint arXiv:1810.00760*, 2018.

D. Binns, E. Dimmer, R. Huntley, D. Barrell, C. O'donovan, and R. Apweiler. Quickgo: a web-based tool for gene ontology searching. *Bioinformatics*, 25(22):3045–3046, 2009.

C. M. Bishop. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1):108–116, 1995.

M. Connor, G. Canal, and C. Rozell. Variational autoencoder with learned latent structure. In *International conference on artificial intelligence and statistics*, pages 2359–2367. PMLR, 2021.

T. R. Davidson, L. Falorsi, N. De Cao, T. Kipf, and J. M. Tomczak. Hyperspherical variational auto-encoders. *arXiv preprint arXiv:1804.00891*, 2018.

L. Falorsi, P. De Haan, T. R. Davidson, N. De Cao, M. Weiler, P. Forré, and T. S. Cohen. Explorations in homeomorphic variational auto-encoding. *arXiv preprint arXiv:1807.04689*, 2018.

D. Grattarola, L. Livi, and C. Alippi. Adversarial autoencoders with constant-curvature latent manifolds. *Applied Soft Computing*, 81:105511, 2019.

D. Kalatzis, D. Eklund, G. Arvanitidis, and S. Hauberg. Variational autoencoders with riemannian brownian motion priors. *arXiv preprint arXiv:2002.05227*, 2020.

D. P. Kingma, M. Welling, et al. Auto-encoding variational bayes, 2013.

M. Kochurov, R. Karimov, and S. Kozlukov. Geoopt: Riemannian optimization in pytorch, 2020.

D. Krioukov, F. Papadopoulos, M. Kitsak, A. Vahdat, and M. Boguná. Hyperbolic geometry of complex networks. *Physical Review E—Statistical, Nonlinear, and Soft Matter Physics*, 82(3):036106, 2010.

F. Locatello, S. Bauer, M. Lucic, G. Raetsch, S. Gelly, B. Schölkopf, and O. Bachem. Challenging common assumptions in the unsupervised learning of disentangled representations. In *international conference on machine learning*, pages 4114–4124. PMLR, 2019.

M. T. Lott, J. N. Leipzig, O. Derbeneva, H. M. Xie, D. Chalkia, M. Sarmady, V. Procaccio, and D. C. Wallace. mtdna variation and analysis using mitomap and mitomaster. *Current protocols in bioinformatics*, 44(1):1–23, 2013.

E. Mathieu, C. Le Lan, C. J. Maddison, R. Tomioka, and Y. W. Teh. Continuous hierarchical representations with poincaré variational auto-encoders. *Advances in neural information processing systems*, 32, 2019.

MITOMAP. Mitomap: A human mitochondrial genome database, 2023. URL `http://www.mitomap.org`. Database.

M. Nickel and D. Kiela. Learning continuous hierarchies in the lorentz model of hyperbolic geometry. In *International conference on machine learning*, pages 3779–3788. PMLR, 2018.

A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

L. A. P. Rey, V. Menkovski, and J. W. Portegies. Diffusion variational autoencoders. *arXiv preprint arXiv:1901.08991*, 2019.

A. Riba. Cell cycle gene regulation dynamics revealed by rna velocity and deep learning, 2021. URL `https://doi.org/10.5281/zenodo.4719436`. Dataset.

A. Riba, A. Oravecz, M. Durik, S. Jiménez, V. Alunni, M. Cerciat, M. Jung, C. Keime, W. M. Keyes, and N. Molina. Cell cycle gene regulation dynamics revealed by rna velocity and deep-learning. *Nature communications*, 13(1):2865, 2022.

S. Schönherr, H. Weissensteiner, F. Kronenberg, and L. Forer. Haplogrep 3-an interactive haplogroup classification and analysis platform. *Nucleic acids research*, 51 (W1):W263–W268, 2023.

V. Schuster and A. Krogh. The deep generative decoder: Map estimation of representations improves modelling of single-cell rna data. *Bioinformatics*, 39(9):btad497, 2023.

O. Skopek, O.-E. Ganea, and G. Bécigneul. Mixed-curvature variational autoencoders. *arXiv preprint arXiv:1911.08411*, 2019.

M. Van Oven. Phylotree build 17: Growing the human mitochondrial dna tree. *Forensic Science International: Genetics Supplement Series*, 5:e392–e394, 2015.

J. Xu and G. Durrett. Spherical latent spaces for stable variational autoencoders. *arXiv preprint arXiv:1808.10805*, 2018.

A. Training details

```
model.z      := init_z(n, manifold) # initialize points on a manifold
model_optim := Adam(model.decoder.parameters())
rep_optim    := RiemannianAdam([model.z])

for each epoch:
    rep_optim.zero_grad()
    for each (i, data) in train_loader:
        model_optim.zero_grad()
        z     := model.z[i]
        z     := add_noise(z, std, manifold) # optional regularization
        y     := model(z)
        loss := loss_fn(y, data)
        loss.backward()
        model_optim.step()
    rep_optim.step()
```

Listing S1: **Pseudocode for training the Riemannian generative decoder.** The necessary changes for learning Riemannian latents rather than Euclidean ones are highlighted with red.

```
def add_noise(z, std, manifold):
    noise     := sample_normal(shape=z.shape) * std
    rie_noise := manifold.egrad2rgrad(x, noise)
    z_noisy   := manifold.retr(x, rie_noise)
    return z_noisy
```

Listing S2: **Pseudocode for adding geometric noise.** `egrad2rgrad` takes a Euclidean gradient and maps it to the Riemannian gradient in the tangent space using the inverse metric. `retr` retracts a tangent vector back onto the manifold via the exponential map if a closed form is available, otherwise a first-order approximation. *geoopt* implements both functions for a wide range of manifolds.

B. Overview of available manifolds

The following are manifolds implemented in *geoopt* (Kochurov et al., 2020), applicable for our representation learning.

- `Euclidean`
- `Stiefel`
- `CanonicalStiefel`
- `EuclideanStiefel`
- `EuclideanStiefelExact`
- `Sphere`
- `SphereExact`
- `Stereographic`
- `StereographicExact`
- `PoincareBall`
- `PoincareBallExact`
- `SphereProjection`
- `SphereProjectionExact`
- `Scaled`
- `ProductManifold`
- `Lorentz`
- `SymmetricPositiveDefinite`
- `UpperHalf`
- `BoundedDomain`

For instance, the torus of Figure 2d is defined using a product manifold of two spheres (i.e., $\mathbb{S}^1 \times \mathbb{S}^1$). Parameterization and further details appear on `geoopt.readthedocs.io`.

C. Experimental details

C.1. Protocols and reproducibility

General details. Non-overlapping train/validation/test splits were made using 82/9/9 percent of samples for each distinct dataset. Across data modalities, our models all use linear layers with hidden sizes $[16, 32, 64, 128, 256]$, and sigmoid linear units (SiLU) as the non-linearity between layers. Decoder parameters were optimized via *Pytorch* (Paszke et al., 2019) with Adam (learning rate 2×10^{-3}, $\beta = (0.9, 0.995)$, weight decay 10^{-3}), a CosineAnnealingWarmRestarts schedule ($T_0 = 40$ epochs) and early-stopped with patience 85, typically resulting in approximately 500 epochs. Representations were optimized with *geoopt*'s RiemannianAdam (learning rate 1×10^{-1}, $\beta = (0.5, 0.7)$, stabilization period 5). Spherical/toroidal representations used learning rate 4×10^{-1} and decoder $\beta = (0.7, 0.9)$. Representations are only updated once an epoch, necessitating larger learning rates and less rigidity via the beta parameters. For hyperbolic manifolds, the curvature was fixed at $c = 5.0$ unless stated otherwise. The cell cycle and branching diffusion models used mean squared error as reconstruction objective, while the hmtDNA models used binary cross entropy.

Test-time RGD latents. Following the strategy of Schuster and Krogh (2023), test-time representations for our model were found by freezing the model parameters and finding optimal z through maximizing the log-likelihood of Equation 5. Note that this biases test reconstruction metrics in favor of our model when compared to encoder-based models.

UMAP parameters. For both the branching diffusion and hmtDNA UMAPs (Figure 3a and Figure 5a), hyperparameters `n_neighbors=30` and `min_dist=0.99` were used to help promote global structure. For Figure 2a, the UMAP coordinates of the original study were used.

Comparison details. We evaluated existing implementations of three baselines: the $\mathcal{P}$-VAE of Mathieu et al. (2019) (based on MIT-licensed `https://github.com/emilemathieu/pvae`), the $\mathcal{S}$-VAE of Davidson et al. (2018) (based on MIT-licensed `https://github.com/nicola-decao/s-vae-pytorch`), and the ΔVAE of Rey et al. (2019) (based on Apache 2.0-licensed `https://github.com/luis-armando-perez-rey/diffusion_vae`). Architectures and hyperparameters were set to best match our model (see the *General details* paragraph) — however for older resources, minor details may differ. Default values were used for any additional parameters.

C.2. Hardware

All experiments were carried out on a Dell PowerEdge R760 server running Linux kernel 4.18.0-553.40.1.el8_10.x86_64. Key components:

- **CPU**: 2 × Intel® Xeon® Gold 6430 (2.1 GHz base, 32 cores/64 threads each; 64 physical cores, 128 threads total)
- **Memory**: 512 GiB DDR5-4800 (8 × 64 GiB RDIMMs)
- **GPUs**: 1 × NVIDIA A30 (24 GB HBM2e; CUDA 12.8; Driver 570.86.15)

Training single-cell and branching diffusion models takes a few minutes on our setup; models on the mitochondrial DNA data train for around 20 minutes.

D. hmtDNA data distributions

Figure S1 shows the distribution of mutation counts for datasets using different reference sequences. Sequence count of each dataset differs since the choice of haplo-tree changes the reported qualities from *Haplogrep3*, affecting the filtering procedure.

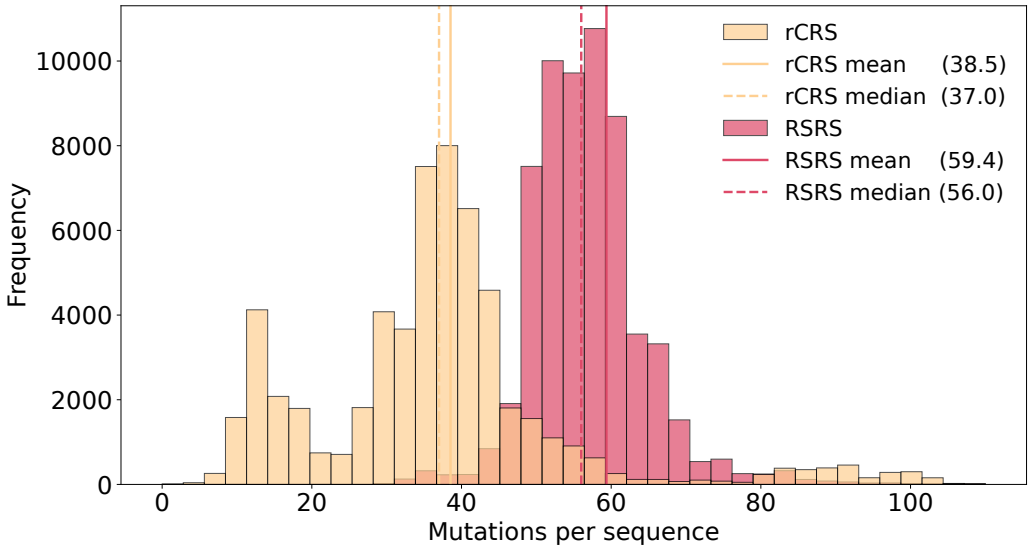


Figure S1: **Distributions of mutation counts** for datasets using different root sequence. Sequence counts of each dataset differ since the choice of haplo-tree changes the reported qualities from *Haplogrep3*, affecting the filtering procedure. Using the revised Cambridge reference sequence (rCRS) means that most sequences contain less mutations when compared to the reconstructed Sapiens reference sequence (RSRS).

E. Geometric regularization: curvature versus noise scale

For a hyperbolic model with curvature c, the metric and its inverse carry a nontrivial, state-dependent factor — which cannot be absorbed into a single global σ.

Concretely, for the ball of curvature c one has

$$G(z) = \frac{4}{(1 - c\,\|z\|^2)^2}\, I, \qquad G^{-1}(z) = \frac{(1 - c\,\|z\|^2)^2}{4}\, I,$$

so the regularizer becomes

$$\mathbb{E}[L(z')] \approx L(z) + \sigma^2\, \frac{(1 - c\,\|z\|^2)^2}{4}\, \mathrm{Tr}\big(J(z)^\top J(z)\big).$$

Changing c thus reshapes the weight $\frac{(1-c\,\|z\|^2)^2}{4}$ across the manifold rather than rescaling a uniform noise-variance. Only at $\|z\| \approx 0$ does it reduce to a constant factor, but in general the curvature and noise-scale contribute distinct effects.

The local noise standard deviation induced by this Riemannian scaling is

$$\sigma(z) = \frac{\sigma\,(1 - c\,\|z\|^2)}{2},$$

which depends on both curvature and position. In particular,

$$\frac{\partial \sigma(z)}{\partial c} = -\frac{\sigma\,\|z\|^2}{2} \;<\; 0 \quad (\|z\| > 0),$$

so increasing c *attenuates* the noise magnitude as one moves away from the origin: If one fixes σ and increases c, then for any $\|z\| > 0$ the factor $(1 - c\,\|z\|^2)$ is smaller, so the actual standard deviation of the injected noise at that point is

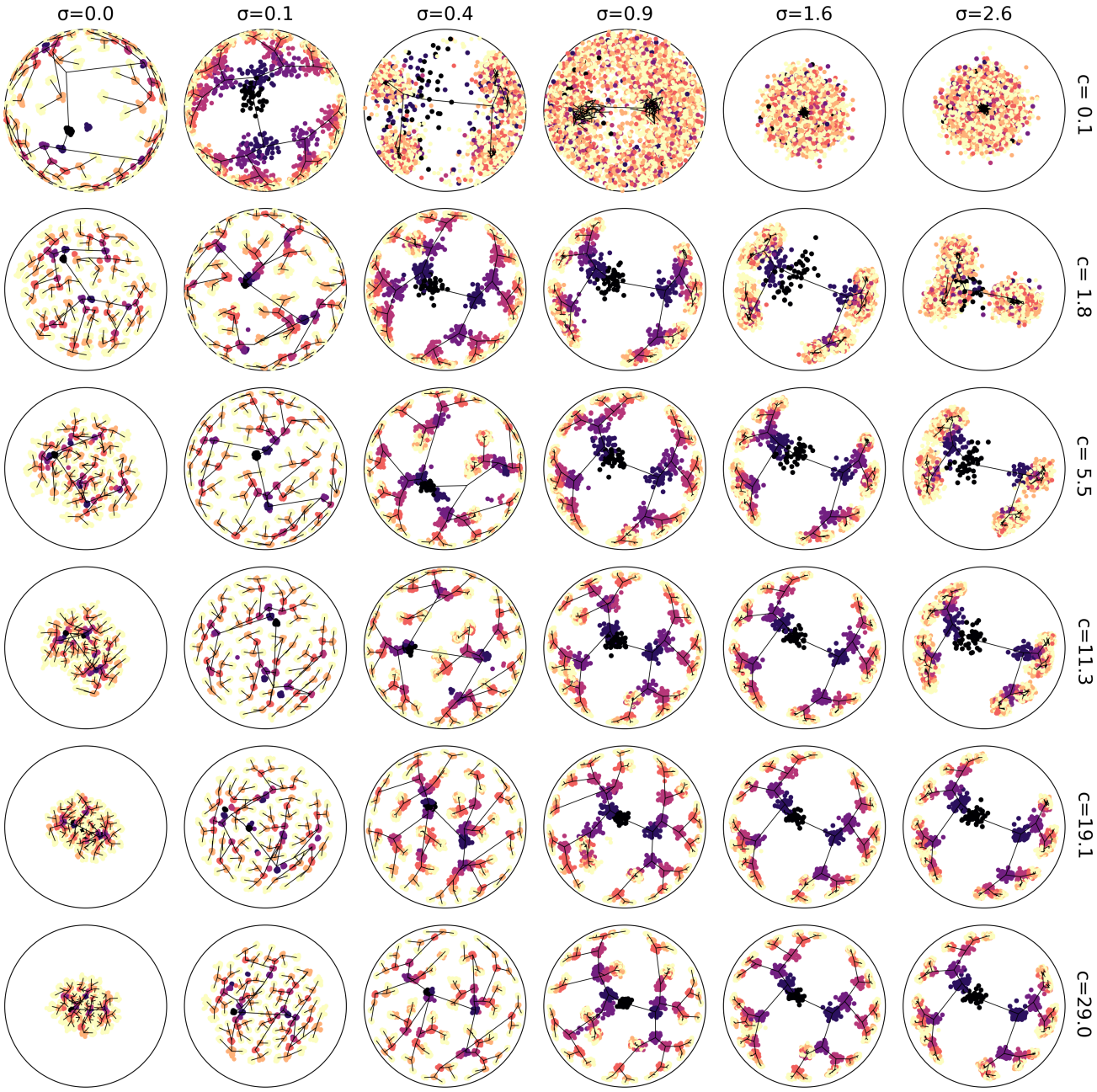


Figure S2: **Effects of manifold curvature and noise level for hyperbolic models on the synthetic branching diffusion dataset**. The visualization is similar to Figure 3c but contains a selection of curvatures rather than $c = 5.0$. Trees consist of 7 levels; color lightness denotes depth.

reduced. Intuitively, points "away from the origin" (larger $\|z\|$) receive less noise. By contrast, raising the global noise scale σ amplifies noise uniformly across all z. Thus curvature c controls the spatial profile of the perturbations, whereas σ governs their overall amplitude. Using the synthetic branching diffusion data, Figure S2 shows the effects of both curvature and noise level.

F. Noise ablation on the hmtDNA sequences

Figure S3 shows the effect of geometry-aware regularization, now on the hmtDNA data.

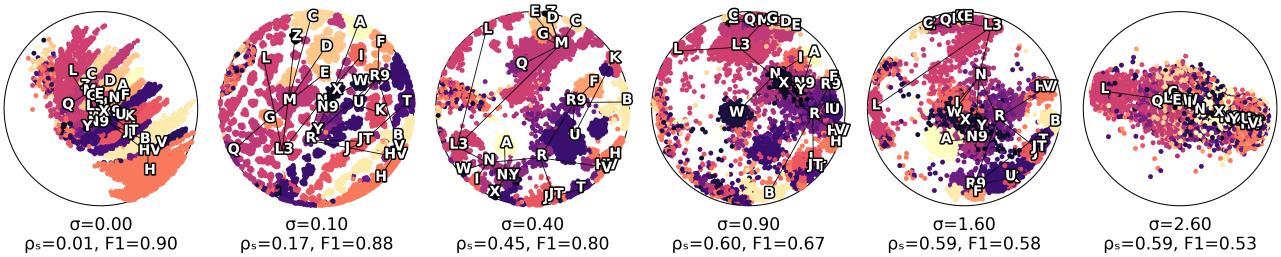


Figure S3: **Gradually increasing σ on the rCRS hmtDNA data**, listing Spearman correlation ρ_s and mean F1-score on the training set. Fixed curvature $c = 5.0$.

G. Errorbar visualization of Table 2

To make Table 2 more digestable, Figure S4 visualizes how the noise level σ impacts correlation and reconstruction metrics for the synthetic branching diffusion dataset.

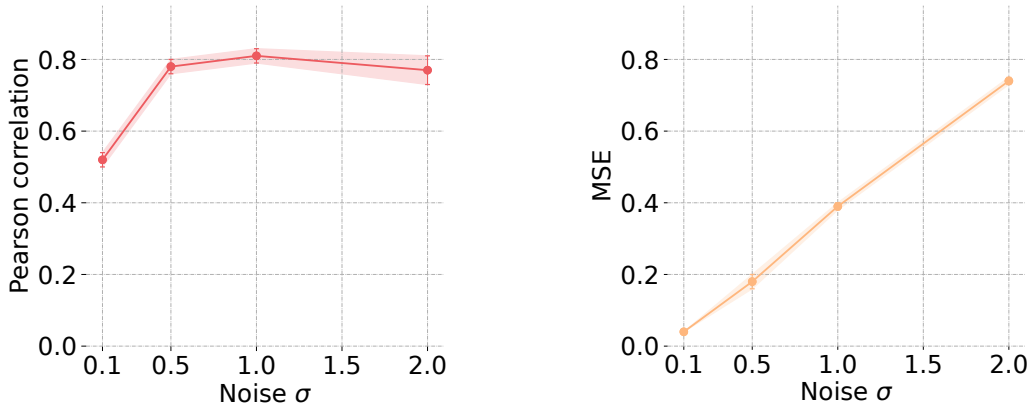


Figure S4: **Errorbar plots varying σ** for the geometry-aware regularization; a visualization of the Lorentz results of Table 2.