Geometry-Aware Autoencoders for Metric Learning and Generative Modeling on Data Manifolds

Xingzhi Sun [*1] **Danqi Liao** [*1] **Kincaid MacDonald** [*1] **Yanlei Zhang** [2 3] **Guillaume Huguet** [2 3] **Guy Wolf** [2 3]
Ian Adelstein [† 4 5] **Tim G. J. Rudner** [† 6] **Smita Krishnaswamy** [† 1 5 7]

Abstract

Non-linear dimensionality reduction methods have proven successful at learning low-dimensional representations of high-dimensional point clouds on or near data manifolds. However, existing methods are not easily extensible— for large datasets, it is prohibitively expensive to add new points to these embeddings. As a result, it is very difficult to use existing embeddings generatively, to sample new points on and along these manifolds. In this paper, we propose GAGA (geometry-aware generative autoencoders) a framework which merges the power of generative deep learning with non-linear manifold learning by: 1) learning generalizable geometry-aware neural network embeddings based on non-linear dimensionality reduction methods like PHATE and diffusion maps, 2) deriving a non-euclidean pullback metric on the data space to generate points faithfully along data manifold geodesics, and 3) learning a flow on the manifold that allows us to transport populations. We provide illustration on easily-interpretable synthetic datasets and showcase results on simulated and real single cell datasets. We show that the geodesic-based generation can be especially important for scientific datasets where the manifold represents a state space and geodesics can represent dynamics of entities over this space.

1. Introduction

There has been rapid growth in high-dimensional scientific data such as scRNA-seq and molecular data. These high-dimensional data are assumed to be concentrated on or near intrinsically low-dimensional manifolds embedded in high-dimensional space. Reliably accessing the intrinsic geometry of the underlying data manifold and computing geometric quantities such as volume, curvature, and geodesics has become increasingly important. For example, in scientific data, the manifold represents a state space, and its geodesics can represent the dynamics of entities over this space.

Non-linear dimensionality reduction methods such as PHATE or diffusion maps have proven useful in learning manifold structure from high-dimensional data. However, they have been difficult to extend to new points or for generative use, to sample new points (Moon et al., 2019; Huguet et al., 2024). To address this, some prior works have tried to regularize an autoencoder to match distances obtained from non-linear dimensionality reduction methods (Huguet et al.; MacDonald et al.; Fasina et al., 2023). Despite distance preservation, these methods have not focused on generative modeling of points, and can struggle in gaps, or sometimes do not decode the data at all and simply provide embeddings (Fasina et al., 2023). As a result, it is very difficult to use existing embeddings for data generation or for sampling new points on and along these manifolds faithfully.

Generating along geodesics on data manifolds often poses a challenge, as it requires a meaningful Riemannian metric on the data manifold to measure arc length and a way to restrict learned paths to stay on the manifold. To tackle this challenge, we propose a generalizable geometry-aware generative autoencoder (GAGA), a neural network-based encoding, which learns an embedding that handles within-manifold points differently from negatively sampled points. First, it embeds within-manifold points in a geometry-preserving fashion by matching distances. Then it folds non-manifold (negatively sampled) points into an auxiliary latent space dimension, embedding off-manifold points far away in latent space. GAGA enables us to extract a continuous metric on the original data space via the Riemannian pullback

[*] Equal contribution . [†] Co-senior authors. [1] Department of Computer Science, Yale University, New Haven CT, USA. [2] Mila - Quebec AI Institute, Montreal, QC, Canada. [3] Department of Mathematics and Statistics, Université de Montréal, QC, Canada. [4] Department of Math, Yale University, New Haven, CT, USA. [5] Applied Math Program, Yale University, New Haven, CT, USA. [6] Center for Data Science, New York University, New York NY, USA. [7] Department of Genetics, Yale University, New Haven, CT, USA.. Correspondence to: Smita Krishnaswamy <smita.krishnaswamy@yale.edu>.

Proceedings of the Geometry-grounded Representation Learning and Generative Modeling at 41st International Conference on Machine Learning, Vienna, Austria. PMLR Vol Number, 2024.

metric, computed using the Jacobian of the neural network encoding. This non-Euclidean pullback metric facilitates data generation along manifold geodesics by measuring the length of curves on the manifold but incurs a sharp penalty in length whenever the curve deviates from the manifold.

Empirically, we show that the tailored geometry-aware embeddings preserve geodesic distances on simulated single-cell datasets, ensuring high DEMaP (Denoised Manifold Affinity Preservation; Moon et al., 2019) consistently under various noise settings. We also demonstrate that GAGA can effectively denoise single-cell data and capture gene-gene correlations. For generative modeling, we show that the induced non-Euclidean metric can generate geodesics on toy datasets and real-world single-cell datasets. Finally, we illustrate that the proposed method can transport distributions through geodesic flows on toy data.

In summary, our main contributions are:

1. We design a geometry-aware generative autoencoder marrying manifold learning with generative modeling.
2. We propose a method for creating a meaningful off-manifold geometry with a non-Euclidean metric on the data space that penalizes movement off the manifold.
3. We present an approach to generating data along geodesics and interpolating between distributions along geodesic flow paths.
4. We demonstrate empirically that the proposed methods work well in practice on toy and real-world data.

2. Background

Manifold Learning and Diffusion Geometry. The *Manifold Hypothesis* states that data often lie *on* or *near* a low-dimensional manifold within high-dimensional space. Manifold learning methods like Diffusion Maps, PHATE, and HeatGeo use diffusion probabilities to recover manifold geometry despite sparsity and noise (Coifman & Lafon, 2006; Moon et al., 2019; Huguet et al., 2024). For details, see Appendix A.

Riemannian Manifolds and Metrics. The *Manifold Hypothesis* encourages using Riemannian geometry tools. An n-dimensional manifold M has a Riemannian metric g for computing angles, lengths, and geodesics. Given a map $f\colon M \to (N, g)$, we induce a geometry on M using the *Riemannian pullback metric*. The differential df of the map pulls back the metric g on N to f^*g on M, defined as $f^*g(X, Y) = g(df_pX, df_pY)$. For details, see Appendix B.

3. Methods

This section is arranged as follows. Section 3.1 presents the distance-matching autoencoder, with an auxiliary dimension informed by a discriminator. This provides a pullback metric on the data space that is matched to the data metric for points on the data manifold, and induces a large distance to the manifold for points off the manifold. Section 3.2 solves the problem of learning geodesics between points on the manifold using this pullback metric. Section 3.3 generalizes this to generating geodesics between populations of points using geodesic-guided flow matching.

3.1. Geometry-Aware Encoding for Both On-Manifold and Off-Manifold Points

Our first task is to learn a latent space embedding whose Euclidean distances correspond to the data manifold geodesic distances. Many existing non-linear dimension reduction techniques, including Diffusion Maps, PHATE, and HeatGeo, achieve this goal. We then produce a non-Euclidean metric on data space that captures data manifold geodesic distances by pulling back (via the encoder) the Euclidean metric from latent space. This metric allows us to generate data along manifold geodesics and interpolate between distributions along geodesic flow paths.

The following standard result from Riemannian geometry states that by matching data manifold geodesic distances in latent space (i.e., learning a local isometry), we construct the desired non-Euclidean pullback metric on the data manifold.

Proposition 1. *For Riemannian manifolds $(\mathcal{M}, g_{\mathcal{M}}), (\mathcal{N}, g_{\mathcal{N}})$ and diffeomorphism $f : \mathcal{M} \to \mathcal{N}$, if f is a local isometry, i.e., there exists $\epsilon > 0$, such that for any $x_0, x_1 \in \mathcal{M}, d_{\mathcal{M}}(x_0, x_1) < \epsilon \implies d_{\mathcal{M}}(x_0, x_1) = d_{\mathcal{N}}(f(x_0), f(x_1))$, then we have $g_{\mathcal{M}} = f^*g_{\mathcal{N}}$.*

To implement this construction, we define an encoder f_θ and a decoder h_ϕ, both parameterized by neural networks, with a reconstruction objective

$$\mathcal{L}_{\text{Recon}}(\theta, \phi) = \frac{1}{N}\sum_{i=1}^{N} ||x_i - h_\phi(f_\theta(x_i))||_2^2, \tag{1}$$

and a distance-matching objective

$$\mathcal{L}_{\text{Dist}}(\theta) \tag{2}$$
$$= \frac{1}{N}\sum_{i<j} e^{-\zeta d(x_i, x_j)} \left(||f_\theta(x_i) - f_\theta(x_j)||_2 - d(x_i, x_j)\right)^2,$$

where $x_1, \dots, x_N$ are the data samples, and $d(x_i, x_j)$ is the geodesic distance between points x_i and x_j obtained via existing dimensionality reduction methods such as PHATE (Moon et al., 2019) or diffusion maps (Coifman & Lafon, 2006). The hyperparameter $\zeta > 0$ and the term $e^{-\zeta d(x_i, x_j)}$ weight the penalty towards the more important local geometry of the data manifold.

To obtain a geometry-aware data encoding, we minimize the objective function

$$\mathcal{L}(\theta, \phi) = \lambda_1 \mathcal{L}_{\text{Dist}}(\theta) + \lambda_2 \mathcal{L}_{\text{Recon}}(\theta, \phi) \tag{3}$$

with respect to θ and ϕ. This objective function balances distance matching and reconstruction with hyperparameters λ_1, λ_2. It results in an embedding that matches the data geometry and retains the information needed to reconstruct the data. The pullback (via the encoder) of the Euclidean metric from latent space yields a non-Euclidean ambient space metric, capturing geodesic distances along the data manifold.

While this construction produces a pullback metric on the entire data space, it is only accurate near the training data, i.e., along the data manifold. To generate data along geodesics without deviating into data space, we create a special embedding for off-manifold points $\check{x}_i$. These points are embedded into a latent space with an auxiliary dimension, far from the embedding of on-manifold points. We achieve this via a GAN-style discriminator that predicts if a point is on or off the manifold and use it to extend the embedding from the previous section.

We first generate negative samples off the data manifold in data space by adding high-dimensional Gaussian noise to the data.

$$\check{x}_i = x_i + \epsilon_i, \;\; \epsilon_i \sim \mathcal{N}(0, cI), \tag{4}$$

where c is a constant chosen such that the space away from the manifold is in the support of the distribution of $\check{x}$. Then, we define a discriminator w_ψ that maps from the ambient space to a score, trained with the Wasserstein-GAN-inspired loss function.

$$\mathcal{L}_w(\psi) = \mathbb{E}_{\check{x}}\left[w_\psi(\check{x})\right] - \mathbb{E}_x\left[w_\psi(x)\right] + \mathrm{Var}_x(w_\psi(x)), \tag{5}$$

w_ψ is a Lipshitz function due to weight clipping and spectral normalization (Arjovsky et al., 2017; Miyato et al., 2018). The variance term is added to encourage the discriminator to have uniform prediction. Finally, we define the off-manifold map

$$r(x) = \begin{pmatrix} f_\theta(x) \\ s(x) \end{pmatrix}, \tag{6}$$

where $s(x) = \beta(\bar{w} - w_\psi(x))$ with $\bar{w} = \mathbb{E}_x[w_\psi(x)]$, and β is a hyperparameter.

For points on the manifold, where $s(x)$ is close to 0, the embedding is unaffected. We formalize this statement as follows:

Lemma 1. *Suppose w_ψ is L-Lipshitz, and $\max_{i,j} ||x_i - \check{x}_j|| \leq M$. for any $\epsilon > 0$, if $\mathcal{L}_w(\psi) \leq -LM + \epsilon$, we have $\mathbb{E}_x[s(x)^2] \leq \epsilon$.*

Points off the manifold, where $s(x)$ is large, are placed into the extended dimension of latent space, far from the on-manifold points. Formally, we have:

Lemma 2. *If there exists $\alpha \in \mathbb{R}$ such that for any $x, \check{x}, \alpha||x - \check{x}|| \leq |w_\psi(x) - w_\psi(\check{x})|$. Then for any $x, \check{x}, ||r(x) - r(\check{x})|| \geq \alpha\beta||x - \check{x}||$. Furthermore, denoting $D_\mathcal{M}(y) := \sup_{x \in \mathcal{M}} ||x - y||$ and $D_{r(\mathcal{M})}(y) := \sup_{x \in \mathcal{M}} ||r(x) - r(y)||$, then for any $\check{x}$, we have $D_{r(\mathcal{M})}(\check{x}) \geq \alpha\beta D_\mathcal{M}(\check{x})$.*

We can now use Equation (6) to extend the pullback metric on the data manifold to the entire ambient space:

Definition 1. The pullback of the Euclidean metric from latent space to the data manifold $\mathcal{M}$ is defined by $g_\mathcal{M}(X, Y) := X^T J_f^T J_f Y$, where $X, Y \in T_x\mathcal{M}$ are tangent vectors at $x \in \mathcal{M}$, $J_f := \partial f_\theta(x)_i / \partial x_j$ is the Jacobian of f_θ at x.

Definition 2. The extended pullback of the Euclidean metric from latent space to the full ambient space $\mathbb{R}^n$ is defined by $g_{\mathbb{R}^n}(X, Y) := X^T J_r^T J_r Y$, where $X, Y \in T_x\mathbb{R}^n$ are tangent vectors at $x \in \mathbb{R}^n$, $J_r := \partial(r(x))_i / \partial x_j$ is the Jacobian of r at x.

Note that $g_\mathcal{M}$ is defined only on the tangent space of $\mathcal{M}$, whereas the off-manifold map $r(x)$ allows $g_{\mathbb{R}^n}$ to be defined on the tangent space of the entire ambient (data) space $\mathbb{R}^n$.

3.2. Learning Geodesics on the Data Manifold

We now turn to the problem of learning the geodesic between a pair of points on the data manifold. One could try to find the curve which minimizes length with respect to the metric $g_\mathcal{M}$. However, this metric is only accurate on the manifold, and such shortest paths might cut through ambient space. Indeed, we need to minimize length under the condition that the curve stays on the manifold. The main result of this section shows that this constrained optimization problem is actually solved by minimizing arc length with respect to the extended pullback $g_{\mathbb{R}^n}$. Intuitively, this metric imposes large penalties for deviating from the manifold, as off-manifold points are embedded into the dimension-extended latent space, forcing the shortest path onto the manifold.

We begin with a neural-network parameterized interpolation curve. for any $x_0, x_1 \in \mathcal{M}$, we define a neural network-parameterized interpolation curve $c_\eta(x_0, x_1, \cdot) : [0, 1] \to \mathbb{R}^n$ satisfying $c_\eta(x_0, x_1, 0) = x_0, c_\eta(x_0, x_1, 1) = x_1$. Further details about the parameterization are provided in Appendix C.1. We minimize

$$\mathcal{L}_{\text{Geo}}(\eta, x_0, x_1) = \frac{1}{M} \sum_{m=1}^{M} g_{\mathbb{R}^n}(\dot{c}_\eta, \dot{c}_\eta)(x_0, x_1, t_m), \tag{7}$$

where $0 = t_0 < t_1 < ... < t_M = 1$ are sampled time points. Note that Equation (7) is a discretization of the integral $\int_0^1 g_{\mathbb{R}^n}(\dot{c}_\eta, \dot{c}_\eta)(x_0, x_1, t)dt$. In Do Carmo & Flaherty Francis (1992), this is defined as the energy of the curve, and following Chapter 9, Proposition 2.5 in Do Carmo & Flaherty Francis (1992), minimizing the energy is equivalent to minimizing the curve length.

The following proposition demonstrates that geodesic computation on $\mathcal{M}$ can be achieved by minimizing arc length with respect to the metric $g_{\mathbb{R}^n}$:

Lemma 3. *Assume that the ω-thickening of $\mathcal{M} \subset \mathbb{R}^n$, $\mathcal{M}^\omega := \{x \in \mathbb{R}^n : \inf_{m \in \mathcal{M}} d(x, m) < \omega\}$, maps into a subset of the ϵ-thickening of $f(\mathcal{M})$, where ϵ can be chosen such that for every $x \in f(\mathcal{M})$, $B_\epsilon \cap f(\mathcal{M})$ has only one connected component. Then, for any smooth $c : [0, 1] \to \mathbb{R}^n$, satisfying $c(0) = x_0, c(1) = x_1$, there exists a smooth $c' : [0, 1] \to \mathcal{M}$, satisfying $c'(0) = x_0, c'(1) = x_1$, such that $\mathcal{L}_{Geo}(c') \leq \mathcal{L}_{Geo}(c) - \alpha^2\beta^2 \frac{1}{M} \sum_{m=1}^{M} (D_{\mathcal{M}}(c(t_m)) - D_{\mathcal{M}}(c(t_{m-1})))^2 + \xi$ where α is in the assumption of Lemma 2 and ξ is a fixed positive constant independent on x_t and β.*

Proposition 2. *When $\mathcal{L}_{Geo}$ is minimized, $\max_{m=1,\ldots,M} D_{\mathcal{M}}(c(t_m)) \leq \sqrt{\xi}/(\alpha\beta)$, i.e., for sufficiently large β, $c(t)$ is close to the manifold with a maximum distance of $\sqrt{\xi}/(\alpha\beta)$. Furthermore, let $c'(t)$ be a geodesic between x_0 and x_1 under the metric $g_{\mathcal{M}}$, we have $\frac{1}{M}\sum_{m=1}^{M} g_{\mathcal{M}}(\dot{c}, \dot{c})(x_0, x_1, t_m) \leq \frac{1}{M}\sum_{m=1}^{M} g_{\mathcal{M}}(\dot{c}', \dot{c}')(x_0, x_1, t_m) + \xi'\sqrt{\xi}/(\alpha\beta)$ for some positive constant ξ'. That is, c approximately minimizes the energy (and hence curve length) under $g_{\mathcal{M}}$.*

This proposition shows that when Equation (7) is minimized, we obtain the geodesic on $\mathcal{M}$ between starting point x_0 and ending point x_1, with respect to the pullback metric $g_{\mathcal{M}}$. We achieve the desired geodesic on $\mathcal{M}$ by minimizing arc length with respect to the extended pullback metric.

3.3. Generation with Geodesic-Guided Flow Matching

In the previous section, we achieved point-wise geodesic computation, learning the geodesic between a pair of points. More generally, we aim to generate population-level geodesics. Given two distributions on the manifold, we want to generate geodesics between populations sampled from these distributions, minimizing the expected total length of the geodesics. This equates to solving the dynamical optimal transport problem (Tong et al., 2020; Benamou & Brenier, 2000), where the cost is the curve length on the manifold.

To solve this, we first find the optimal pairing of points from the starting and ending distributions to minimize total geodesic length, then compute those geodesics. To generalize to new points, we learn a vector field matching the time derivatives (speed) of the geodesics. Given a point sampled from the first distribution, we can generate the geodesic by integrating the vector field starting from the point.

Specifically, we define a neural network $v_\nu(x_0, t) \in \mathbb{R}^n$, and the flow matching loss for any joint distribution π and curve c_η:

$$\mathcal{L}_{\mathrm{FM}}(\nu, \eta, x_0, x_1) = \mathbb{E}_{\pi(x_0,x_1)} ||v_\nu(t, x_0) - \frac{d}{dt} c_\eta(t, x_0, x_1)||^2. \tag{8}$$

When this loss is minimized, v_ν is the vector field that matches the time derivatives of the curves.

In each training step, we sample starting and ending points from the two distributions, and solve the optimal transport problem where the ground distance is the Euclidean distance in the latent space. This optimal transport plan π would minimize the total geodesic length between $(x_0, x_1) \sim \pi$, because GAGA is trained so that the Euclidean distance in the latent space is matched to the geodesic distance on the data manifold. We then parameterize the interpolation curves c_η as in Section 3.2, and minimize the following loss which balances the loss Equation (7) that the c_η are the geodesics, and the aforementioned flow matching lossEquation (8).

$$\mathcal{L}_{\mathrm{GFM}}(\nu, \eta, x_0, x_1) = \lambda_3 \mathcal{L}_{\mathrm{geo}}(\eta, x_0, x_1) + \lambda_4 \mathcal{L}_{\mathrm{FM}}(\nu, \eta, x_0, x_1). \tag{9}$$

Further details are provided in Algorithm 1.

After training, we generate the geodesics by integrating the vector field v_η. Given an initial point x_0, we can generate points along the geodesics starting from it with

$$x(t) = x_0 + \int_0^t v_\nu(x_0, \tau) d\tau. \tag{10}$$

Finally, we have the following proposition that shows our method generates the desired population-level geodesics:

Proposition 3. *Given starting and ending distributions p, q, at the convergence of Algorithm 1,*

$$x(t) = x_0 + \int_0^t v_\nu(x_0, \tau) d\tau, x_0 \sim p, t \in [0, 1] \tag{11}$$

are geodesics between the two distributions following the optimal transport plan that minimizes the total expected geodesic lengths.

4. Results

Geometry-aware autoencoder. We empirically show that GAGA preserves geodesic distances of data manifold in the latent space and can effectively recover gene trends through decoding. We evaluate GAGA on synthetic scRNA-seq datasets Splatter (Zappia et al., 2017).

For the encoder, we use DEMaP (Denoised Embedding Manifold Preservation; Moon et al., 2019) to measure the

Table 1. Average DEMaP and DRS on simulated single-cell datasets over different noise settings.

	Cellular State Space	DEMaP (↑)	DRS (↑)
Autoencoder	Clusters	0.347±0.117	0.642±0.129
GAGA	Clusters	**0.645**±0.195	**0.667**±0.165
Autoencoder	Trajectories	0.433±0.135	**0.587**±0.148
GAGA	Trajectories	**0.600**±0.191	0.559±0.143

Table 2. Average MSE between predicted geodesic lengths and ground truth on data with different dimensions and noise settings.

	Djikstra's	GAGA	Local	Density
Ellipsoid	4.40±6.6	**3.76**±7.1	143.70±246.5	5.36±9.0
Hemisphere	4.83±6.2	**0.47**±0.6	43.20±65.7	0.60±0.6
Saddle	**1.87**±3.5	4.11±8.8	55.59±76.8	5.30±12.4
Torus	5.01±7.9	**4.09**±6.3	271.84±295.3	5.39±9.5

correlation between Euclidean distances in latent space and ground truth geodesic distances in data space. For decoder, we propose a novel criterion, DRS (Denoised Reconstruction Score) to compute the correlation between reconstructed genes and denoised genes through denoising method MAGIC (Van Dijk et al., 2018). Every set of our experiments is repeated under 5 random seeds. See Appendix E.1 for details on Splatter and evaluation criteria.

We can see from Table 1 that the distance matching loss is important for preserving geodesic distances, indicated by higher DEMaP scores averaged across different noise levels. Furthermore, GAGA generally rivals the standard autoencoder on DRS, indicating our distance-matching loss does not degrade data reconstruction.

Generating along geodesics on the data manifold. We use synthetic manifold datasets to evaluate generated geodesics. We compare our method with Dijkstra's algorithm, a baseline that directly uses the local metric, and a method that uses density regularization. Further details are provided in Appendix E.2.

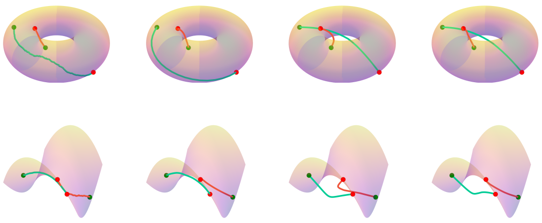

Figure 1. Comparison of geodesics. From left to right columns: 1) ground truth, 2) GAGA, 3) local metric, 4) density regularization.

Table 2 shows that GAGA generally outperforms all the other methods except Djikstra's on the saddle datasets. It's noteworthy to point out that Dijkstra's algorithm is only capable of connecting existing points but is unable to generate points along the path. Directly using the local metric performs the worst, lagging far behind all other methods.

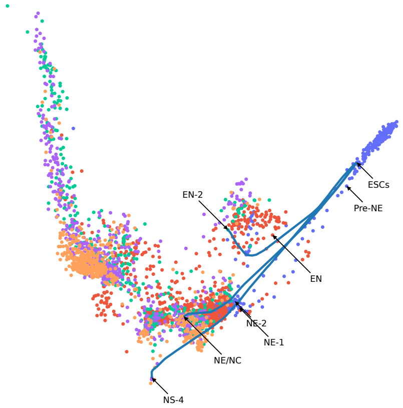

Figure 2. Geodesics learned for the Embryoid Body dataset.

We visualize the predicted geodesics in Figure 1. In general, trajectories generated by GAGA stay on the manifold and are close to the ground truth geodesics, whereas some learned by the local metric or density regularization either deviate from the ground truth or cut through the manifold. Further details are provided in Appendix F.3.

In addition to toy datasets, we also visualize the geodesics learned on the Embryoid Body dataset (Figure 2). The starting points correspond to stem cells, while the ending points are selected at different lineages. The predicted geodesics recover the corresponding differentiation branches, aligning with our biological understanding of the data.

Geodesics-guided flow matching. Lastly, we evaluate geodesics-guided flow matching on several toy datasets. Figure 3 shows that the starting population of points is successfully transformed into the ending population through geodesic flows on the manifold.

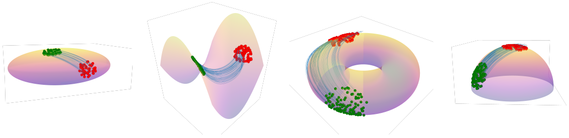

Figure 3. Generating transporting trajectories on toy manifolds.

5. Conclusion

In this paper, we propose a geometric-aware generative autoencoder (GAGA) that preserves geometry in latent embeddings and can generate new points on the data manifold, along the geodesics, and at the population levels. We circumvent the limitations of existing manifold learning methods by training generalizable geometric-aware neural network embeddings, folding off-manifold points into auxiliary latent dimension, and learning a non-euclidean metric on data space via Riemannian pullback metric.

Acknowledgements

This research was partially funded and supported by ESP *Mérite* [G.H.], CIFAR AI Chair [G.W.], NSERC Discovery grant 03267 [G.W.], NIH grants (1F30AI157270-01, R01HD100035, R01GM130847, R01GM135929) [G.W.,S.K.], NSF Career grant 2047856 [S.K.], NSF grant 2327211 [S.K.,G.W.,M.P.,I.A.], NSF/NIH grant 1R01GM135929-01 [M.H.,S.K.], the Chan-Zuckerberg Initiative grants CZF2019-182702 and CZF2019-002440 [S.K.], the Sloan Fellowship FG-2021-15883 [S.K.], and the Novo Nordisk grant GR112933 [S.K.]. The content provided here is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies. The funders had no role in study design, data collection & analysis, decision to publish, or preparation of the manuscript.

References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *International conference on machine learning*, pp. 214–223. PMLR, 2017.

Benamou, J.-D. and Brenier, Y. A computational fluid mechanics solution to the monge-kantorovich mass transfer problem. *Numerische Mathematik*, 84(3):375–393, 2000.

Coifman, R. R. and Lafon, S. Diffusion maps. *Applied and computational harmonic analysis*, 21(1):5–30, 2006.

Do Carmo, M. P. and Flaherty Francis, J. *Riemannian geometry*, volume 2. Springer, 1992.

Fasina, O., Huguet, G., Tong, A., Zhang, Y., Wolf, G., Nickel, M., Adelstein, I., and Krishnaswamy, S. Neural fim for learning fisher information metrics from point cloud data. In *International Conference on Machine Learning*, pp. 9814–9826. PMLR, 2023.

Huguet, G., Magruder, D. S., Tong, A., Fasina, O., Kuchroo, M., Wolf, G., and Krishnaswamy, S. Manifold Interpolating Optimal-Transport Flows for Trajectory Inference. URL http://arxiv.org/abs/2206.14928.

Huguet, G., Tong, A., De Brouwer, E., Zhang, Y., Wolf, G., Adelstein, I., and Krishnaswamy, S. A heat diffusion perspective on geodesic preserving dimensionality reduction. *Advances in Neural Information Processing Systems*, 36, 2024.

MacDonald, K., Bhaskar, D., Thampakkul, G., Nguyen, N., Zhang, J., Perlmutter, M., Adelstein, I., and Krishnaswamy, S. A Flow Artist for High-Dimensional Cellular Data. In *2023 IEEE 33rd International Workshop on Machine Learning for Signal Processing (MLSP)*, pp. 1–6. doi: 10.1109/MLSP55844.2023.10285942. URL https://ieeexplore.ieee.org/document/10285942.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Moon, K. R., Van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., Yim, K., Elzen, A. v. d., Hirn, M. J., Coifman, R. R., et al. Visualizing structure and transitions in high-dimensional biological data. *Nature biotechnology*, 37(12):1482–1492, 2019.

Tong, A., Huang, J., Wolf, G., Van Dijk, D., and Krishnaswamy, S. Trajectorynet: A dynamic optimal transport network for modeling cellular dynamics. In *International conference on machine learning*, pp. 9526–9536. PMLR, 2020.

Tong, A., Malkin, N., Huguet, G., Zhang, Y., Rector-Brooks, J., Fatras, K., Wolf, G., and Bengio, Y. Improving and generalizing flow-based generative models with minibatch optimal transport. *arXiv preprint arXiv:2302.00482*, 2023.

Van Dijk, D., Sharma, R., Nainys, J., Yim, K., Kathail, P., Carr, A. J., Burdziak, C., Moon, K. R., Chaffer, C. L., Pattabiraman, D., et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 174(3): 716–729, 2018.

Zappia, L., Phipson, B., and Oshlack, A. Splatter: simulation of single-cell rna sequencing data. *Genome biology*, 18(1):174, 2017.

Appendix

A. Manifold Learning and Diffusion Geometry

The *Manifold Hypothesis* states that data are often sampled *on* or *near* an intrinsically low-dimensional manifold within high-dimensional Euclidean space. Manifold learning techniques aim to uncover and recreate this manifold in a lower-dimensional space.

Many manifold learning approaches use data *diffusion geometry*, which extracts geometric features from an approximation of heat flow on the data. Diffusion geometry models a high-dimensional point cloud as a graph by applying a kernel $\mathcal{K}$ (e.g., the Gaussian kernel $\exp(-||z_1 - z_2||^2/\sigma)$) to the pairwise Euclidean distances between data points.

The kernel $\mathcal{K}$ is normalized to obtain a row-stochastic matrix P, where $P(z_1, z_2) = \frac{\mathcal{K}(z_1,z_2)}{||\mathcal{K}(z_1,\cdot)||_1}$. This matrix P encodes transition probabilities between points. Powering P^t represents a t-step random walk. Long-range or spurious connections are given less weight through this iterated walk than robust on-manifold paths, allowing the resulting point-wise *diffusion probabilities* to recover manifold geometry even in the presence of sparsity and noise. Methods like Diffusion Maps, PHATE, and HeatGeo use diffusion probabilities to define a *statistical distance* between data points (Coifman & Lafon, 2006; Moon et al., 2019; Huguet et al., 2024).

B. Riemannian Manifolds & Metrics

The *Manifold Hypothesis* also encourages the use of tools from Riemannian geometry. Formally, an n-dimensional manifold M is a topological space locally homeomorphic to n-dimensional Euclidean space. Riemannian manifolds (M, g) have the additional structure of a Riemannian metric g, an inner product defined on the tangent space of each point, enabling the computation of angles, lengths, and geodesics.

Given a map between manifolds $f\colon M \to (N, g)$, where the target manifold (N, g) has a Riemannian metric, we can induce a geometry on the original manifold M through the *Riemannian pullback metric*. First, we define the differential df of the map, which at a point $p \in M$, is a map between tangent spaces $df_p\colon T_pM \to T_{f(p)}N$. We use this map df to pull back the metric g on N to a metric f^*g on M.

To define this pullback, we specify its effect on a pair of vectors $X, Y \in T_pM$. The pullback metric is defined as

$$f^*g(X, Y) = g(df_pX, df_pY), \tag{B.1}$$

where df_pX and df_pY are the pushforward vectors, i.e., the images of X and Y under the differential.

C. Geodesic Parameterization and Computation

C.1. Parameterization of curve

We parameterize the curves using an interpolation between starting and ending points, with a linear term and a non-linear term parameterized by an MLP γ_η.

$$c_\eta(x_0, x_1, t) = tx_1 + (1-t)x_0 + (1-(2t-1)^2)\gamma_\eta(x_0, x_1, t), \tag{C.2}$$

C.2. Algrorithm for Geodesic Flow Matching

We use geodesic flow matching with a minibatch OT setup similar to (Tong et al., 2023)

D. Proofs

D.1. Proposition 1

For Riemannian manifolds $(\mathcal{M}, g_\mathcal{M}), (\mathcal{N}, g_\mathcal{N})$ and diffeomorphism $f : \mathcal{M} \to \mathcal{N}$, if f is a local isometry, i.e., there exists $\epsilon > 0$, such that for any $x_0, x_1 \in \mathcal{M}, d_\mathcal{M}(x_0, x_1) < \epsilon \implies d_\mathcal{M}(x_0, x_1) = d_\mathcal{N}(f(x_0), f(x_1))$, then we have $g_\mathcal{M} = f^*g_\mathcal{N}$.

Proof. We first prove that the two metrics agree on vector norms. That is, for any $u \in T_x\mathcal{M}, g_\mathcal{N}(dfu, dfu) = g_\mathcal{M}(u, u)$.:

Algorithm 1 Mini-batch OT Geodesic Flow Matching

```
Input: Starting and ending populations $\mathcal{X}, \mathcal{Y}$, encoder $f$, dimension-extended encoder $r$, $t = (t_1, ..., t_M)$
while Training do
  Sample batches of size $b$ i.i.d. from the datasets
  Sample $\{x_1, ..., x_l\} \subset \mathcal{X}, \{y_1, ..., y_l\} \subset \mathcal{Y}$
  $\mu \leftarrow \frac{1}{l}\sum_{i=1}^{l} I(x = x_i), \nu \leftarrow \frac{1}{l}\sum_{i=1}^{l} I(x = y_i)$
  $\pi^* = \underset{\pi \sim \Gamma(\mu,\nu)}{\arg\min} \left(\frac{1}{l}\sum_{i=1}^{l} \pi(x'_i, y'_i)||f(x'_i) - f(y'_i)||^2\right)^{1/2}$
  Sample $(x_{j_1}, y_{j_1}), ..., (x_{j_l}, y_{j_l}) \overset{i.i.d.}{\sim} \pi^*$
  Compute geodesic and velocity-matching losses
  $L \leftarrow \frac{1}{l}\sum_{i=1}^{l}(\lambda_3\mathcal{L}_{\text{geo}}(\eta, x_{j_i}, y_{j_i}) + \lambda_4\mathcal{L}_{\text{FM}}(\nu, \eta, x_{j_i}, y_{j_i}))$
  $\eta, \nu \leftarrow \mathrm{GradientDescentUpdate}(\eta, \nu, \nabla L)$
end while
Output: $\nu$
```

$\forall z \in \mathcal{N}, \forall$ smooth curve $\gamma(t) \subset \mathcal{N}$, and let $\xi(t) = f^{-1}(\gamma(t))$. Then there exists $\delta > 0$ such that $\forall 0 < t < \delta$

$$\int_0^t \sqrt{g_{\mathcal{M}}(\dot{\xi}(\tau), \dot{\xi}(\tau))}d\tau < \epsilon \tag{D.3}$$

We have

$$\int_0^t \sqrt{g_{\mathcal{M}}(\dot{\xi}(\tau), \dot{\xi}(\tau))}d\tau = \int_0^{\gamma^{-1}\circ\xi(t)} \sqrt{g_{\mathcal{N}}(\dot{\gamma}(\tau), \dot{\gamma}(\tau))}d\tau \tag{D.4}$$

Take $t \to 0$, we have $g_{\mathcal{N}}(dfu, dfu) = g_{\mathcal{M}}(u, u)$ where $u = \dot{\xi}(0)$.

Next we use the identity

$$\langle u, v\rangle = \frac{1}{4}\left(\langle u + v, u + v\rangle - \langle u - v, u - v\rangle\right) \tag{D.5}$$

for any 2-form $\langle\cdot,\cdot\rangle$, and apply to $g_{\mathcal{M}}, g_{\mathcal{N}}$, we have

$$g_{\mathcal{N}}(dfu, dfv) = g_{\mathcal{M}}(u, v)\forall u, v \in T_x\mathcal{M}. \tag{D.6}$$

□

D.2. Lemma 1

Suppose w_ψ is L-Lipshitz, and $\max_{i,j}||x_i - \breve{x}_j|| \le M$. $\forall\epsilon > 0$, if $\mathcal{L}_w(\psi) \le -LM + \epsilon$, we have $\mathbb{E}_x[s(x)^2] \le \epsilon$.

Proof. Denote p_{on} the data distribution and p_{off} the distribution of off-manifold points defined Equation (4).

$\forall x \sim p_{\text{on}}, \breve{x} \sim p_{\text{off}}$, since w_ψ is L-Lipshitz, $|w_\psi(\breve{x}) - w_\psi(x)| \le L||\breve{x} - x|| < LM$.

Taking expectaion, we have $\mathbb{E}_{\breve{x}}[w_\psi(\breve{x})] - E_x[w_\psi(x)] \ge -LM$.

Thus, $\mathcal{L}_w(\psi) \le -LM + \epsilon \implies \mathbb{E}[s(x)^2] = \mathrm{Var}_x(w_\psi(x)) = \mathcal{L}_w(\psi) - (\mathbb{E}_{\breve{x}}[w_\psi(\breve{x})] - E_x[w_\psi(x)]) \le \epsilon$. □

D.3. Lemma 2

If there exists $\alpha \in \mathbb{R}$ such that for any $x, \breve{x}, \alpha||x - \breve{x}|| \le |w_\psi(x) - w_\psi(\breve{x})|$. Then for any $x, \breve{x}, ||r(x) - r(\breve{x})|| \ge \alpha\beta||x - \breve{x}||$. Furthermore, denoting $D_{\mathcal{M}}(y) := \sup_{x\in\mathcal{M}}||x - y||$ and $D_{r(\mathcal{M})}(y) := \sup_{x\in\mathcal{M}}||r(x) - r(y)||$, then for any $\breve{x}$, we have $D_{r(\mathcal{M})}(\breve{x}) \ge \alpha\beta D_{\mathcal{M}}(\breve{x})$.

Proof. Because $r(x) = \begin{pmatrix} f_\theta(x) \\ s(x) \end{pmatrix}$, where $s(x) = \beta(\bar{w} - w_\psi(x))$, we directly compute:

$$||r(x) - r(\check{x})||^2 = ||f_\theta(x) - f_\theta(\check{x})||^2 + |s(x) - s(\check{x})|^2 \tag{D.7}$$

$$\geq |s(x) - s(\check{x})|^2 \tag{D.8}$$

$$\geq \beta^2 |w_\psi(x) - w_\psi(\check{x})|^2 \tag{D.9}$$

$$\geq \beta^2 \alpha^2 ||x - \check{x}||^2, \tag{D.10}$$

we have $||r(x) - r(\check{x})|| \geq \beta\alpha ||x - \check{x}||$.

Taking supremum over $x \in \mathcal{M}$, we have $D_{r(\mathcal{M})}(\check{x}) \geq \beta\alpha D_{\mathcal{M}}(\check{x})$ □

D.4. Lemma 3

Assume that the ω-thickening of $\mathcal{M} \subset \mathbb{R}^n$, $\mathcal{M}^\omega := \{x \in \mathbb{R}^n : \inf_{m \in \mathcal{M}} d(x, m) < \omega\}$, maps into a subset of the ϵ-thickening of $f(\mathcal{M})$, where ϵ can be chosen such that for every $x \in f(\mathcal{M})$, $B_\epsilon \cap f(\mathcal{M})$ has only one connected component. Then, for any smooth $c : [0,1] \to \mathbb{R}^n$, satisfying $c(0) = x_0, c(1) = x_1$, there exists a smooth $c' : [0,1] \to \mathcal{M}$, satisfying $c'(0) = x_0, c'(1) = x_1$, such that $\mathcal{L}_{\text{Geo}}(c') \leq \mathcal{L}_{\text{Geo}}(c) - \alpha^2\beta^2 \frac{1}{M}\sum_{m=1}^{M}(D_{\mathcal{M}}(c(t_m)) - D_{\mathcal{M}}(c(t_{m-1})))^2 + \xi$ where α is in the assumption of Lemma 2 and ξ is a fixed positive constant independent on x_t and β.

Proof. Consider a smooth $c : [0,1] \to \mathbb{R}^n$ with $c(0) = x_1, c(0) = x_1$ which lies within the ω-thickening of $\mathcal{M}$. We construct an open cover of its image $f(c)$ as the collection of open balls $\{B_\epsilon(c(t)) : t \in [0,1]\}$. By compactness, this admits a finite subcover at some collection of times $\{t_1 \ldots t_N\}$. For each t_i, we can choose point $c'[t_i]$ from $B_\epsilon(c(t_i)) \cap f(\mathcal{M})$. By the continuity of $f \circ c$, these are all part of the same connected component of $f(\mathcal{M})$, hence there exists a curve $c' : [0,1] \to \mathbb{R}^n$ with the same endpoints as c, whose image contains $\{c'[t_i]\}$. Furthermore, by the smoothness of f and c, there exists a uniform $K > 0$ independent of c, c' such that $|\int \dot{c}(t)^T J_f^T J_f \dot{c}(t) - \dot{c}'(t)^T J_f^T J_f \dot{c}'(t) dt| < K\epsilon$. Following Lemma 1, because $c' \in \mathcal{M}$, we also have $|\int \dot{c}'(t)^T J_s^T J_s \dot{c}'(t)| < \epsilon'$ for some uniform $\epsilon' > 0$ independent on c, c'.

We can decompose the pullback metric as

$$J_r^T J_r = J_f^T J_f + J_s^T J_s. \tag{D.11}$$

and compute the difference

$$\mathcal{L}_{\text{Geo}}(c) - \mathcal{L}_{\text{Geo}}(c') = \frac{1}{M}\sum_{m=1}^{M}(\dot{c}(t)^T J_f^T J_f \dot{c}(t) + \dot{c}(t)^T J_s^T J_s \dot{c}(t) - (\dot{c}'(t)^T J_f^T J_f \dot{c}'(t) + \dot{c}'(t)^T J_s^T J_s \dot{c}'(t))) \tag{D.12}$$

$$= \frac{1}{M}\sum_{m=1}^{M}(\dot{c}(t)^T J_f^T J_f \dot{c}(t) - \dot{c}'(t)^T J_f^T J_f \dot{c}'(t) + \dot{c}(t)^T J_s^T J_s \dot{c}(t) + \dot{c}'(t)^T J_s^T J_s \dot{c}'(t)) \tag{D.13}$$

$$\geq -K\epsilon - \epsilon' + \frac{1}{M}\sum_{m=1}^{M}\dot{c}(t)^T J_s^T J_s \dot{c}(t). \tag{D.14}$$

$$\geq -K\epsilon - \epsilon' - \epsilon'' + \frac{1}{M}\sum_{m=1}^{M}\dot{(}s(c(t_m)) - s(c(t_{m-1}))^2 \tag{D.15}$$

$$\geq -K\epsilon - \epsilon' - \epsilon'' + \frac{1}{M}\sum_{m=1}^{M}\dot{(}s(c(t_m)) - s(c(t_{m-1}))^2. \tag{D.16}$$

$$\geq -K\epsilon - \epsilon' - \epsilon'' + \frac{1}{M}\alpha\beta\sum_{m=1}^{M}\dot{(}D_{\mathcal{M}}(c(t_m)) - D_{\mathcal{M}}(c(t_{m-1}))^2, \tag{D.17}$$

$$\tag{D.18}$$

where ϵ', ϵ'' are positive constants independent on x_t, β. □

D.5. Proposition 2

When $\mathcal{L}_{\text{Geo}}$ is minimized, $\max\limits_{m=1,\dots,M} D_{\mathcal{M}}(c(t_m)) \leq \frac{\sqrt{\xi}}{\alpha\beta}$, i.e., for sufficiently large β, $c(t)$ is close to the manifold with a maximum distance of $\frac{\sqrt{\xi}}{\alpha\beta}$. Furthermore, let $c'(t)$ be a geodesic between x_0 and x_1 under the metric $g_{\mathcal{M}}$, we have $\frac{1}{M}\sum_{m=1}^{M} g_{\mathcal{M}}(\dot{c},\dot{c})(x_0,x_1,t_m) \leq \frac{1}{M}\sum_{m=1}^{M} g_{\mathcal{M}}(\dot{c}',\dot{c}')(x_0,x_1,t_m) + \xi' \frac{\sqrt{\xi}}{\alpha\beta}$ for some positive constant ξ'. That is, c approximately minimizes the energy (and hence curve length) under $g_{\mathcal{M}}$.

Proof. Suppose c minimizes $\mathcal{L}_{\text{Geo}}$. Then by Lemma 3, there exists c' such that

$$\mathcal{L}_{\text{Geo}}(c') \leq \mathcal{L}_{\text{Geo}}(c) - \alpha^2\beta^2 \frac{1}{M}\sum_{m=1}^{M}(D_{\mathcal{M}}(c(t_m)) - D_{\mathcal{M}}(c(t_{m-1})))^2 + \xi. \tag{D.19}$$

On the other hand, because c is a minimizer, we have

$$\mathcal{L}_{\text{Geo}}(c) \leq \mathcal{L}_{\text{Geo}}(c'). \tag{D.20}$$

Combining them, we have

$$\alpha^2\beta^2 \frac{1}{M}\sum_{m=1}^{M}(D_{\mathcal{M}}(c(t_m)) - D_{\mathcal{M}}(c(t_{m-1})))^2 \leq \mathcal{L}_{\text{Geo}}(c) - \mathcal{L}_{\text{Geo}}(c') + \xi \leq \xi. \tag{D.21}$$

Rearrange $t_0, \dots, t_M$ with a permutation σ such that $D_{\mathcal{M}}(t_{\sigma(0)}) \leq \cdots \leq D_{\mathcal{M}}(t_{\sigma(M)})$, and because $D_{\mathcal{M}}(t_0) = 0$ (the minimum), WLOG, let $t_{\sigma(0)} = 0$. We have

$$\begin{aligned}
\alpha^2\beta^2 \frac{1}{M}\sum_{m=1}^{M}(D_{\mathcal{M}}(c(t_{\sigma(m)})) - D_{\mathcal{M}}(c(t_{\sigma(m-1)})))^2 &\leq \xi && \text{(D.22)}\\
\implies \alpha^2\beta^2 (\frac{1}{M}\sum_{m=1}^{M}(D_{\mathcal{M}}(c(t_{\sigma(m)})) - D_{\mathcal{M}}(c(t_{\sigma(m-1)})))^2 &\leq \xi \text{ (by Jensen's inequality)} && \text{(D.23)}\\
\implies \alpha^2\beta^2 (D_{\mathcal{M}}(c(t_{\sigma(M)}) - D_{\mathcal{M}}(c(t_{\sigma(0)}))^2 &\leq \xi && \text{(D.24)}\\
\implies \max_{m=1,\dots,M} D_{\mathcal{M}}(c(t_m)) = D_{\mathcal{M}}(c(t_{\sigma(M)})) \leq \frac{\sqrt{\xi}}{\alpha\beta}. & && \text{(D.25)}
\end{aligned}$$

The proof for the second part follows from the Lipshitz property of $s(x)$ and the smoothness of f in Lemma 3. □

D.6. Proposition 3

At the convergence of Algorithm 1, Equation (10) are geodesics between points in $\mathcal{X}$ and points in $\mathcal{Y}$ following the optimal transport plan that minimizes the geodesic lengths.

Proof. We first prove that when Equation (7) and Equation (8) are minimized, Equation (10) yields geodesics from $x_0 \in \mathcal{X}$ to $x_1 \in \mathcal{Y}$. This is because by Lemma 3, the curves c_η are geodesics. When Equation (8) is minimized, v_ν approximates the gradient of c_η, and its integration starts at the same point x_0 approximates c_η.

The rest follows from the the proof of Algorithm 3 in (Tong et al., 2023).

□

E. Experiment Details

E.1. Geometry-aware autoencoder

E.1.1. Datasets: Splatter

We evaluate our geometry-aware autoencoder on simulated scRNA-seq datasets Splatter(Zappia et al., 2017). Splatter uses parametric models to simulate cell populations with multiple cell types, structures, and differentiation patterns. Specifically, we evaluate on single-cell data of group and path structures with biological coefficient of variation (bcv) parameters $\{0, 0.18, 0.25, 0.5\}$. A higher bcv corresponds to a lower signal-to-noise ratio. The cellular state space is a simulation parameter indicating whether the cells are arranged in clusters or trajectories in the data space. In Splatter, it is specified by the `method` parameter, where clusters correspond to `groups` and trajectories correspond to `paths`.

E.1.2. Evaluation Criteria

For the encoder, we leverage DEMaP (Moon et al., 2019) to measure the correlation between Euclidean distances in latent space and ground truth geodesic distances in original data space.

$$\text{DEMaP}(f) = \frac{2}{N(N-1)} \sum_{i<j} \mathrm{Corr}(||f(x_i) - f(x_j)||_2, d_{ij}), \tag{E.26}$$

where f is the encoder to be evaluated, Corr is Pearson correlation, x_i, x_j are points from test data, and d_{ij} is the ground truth geodesic distance between x_i, x_j, computed from shortest path distance under noiseless setting.

For decoder evaluation, we propose a novel criteria, DRS (Denoised Reconstruction Score), to account for the noisy and sparse nature of single-cell data. DRS computes the correlation between reconstructed genes and denoised genes through denoising and imputation method MAGIC(Van Dijk et al., 2018).

$$\text{DRS}(f, h) = \frac{1}{N_{\text{gene}}} \sum_{i=1}^{N_{\text{gene}}} \mathrm{Corr}(y_i, y_i^{\text{MAGIC}}), \tag{E.27}$$

where f, h are the encoder and decoder pair, $y_i = \mathrm{PCA}^{-1}(h(f(x_i)))$, $y_i^{\text{MAGIC}} = \mathrm{PCA}^{-1}(\mathrm{MAGIC}(x_i))$. PCA^{-1} here is the inverse PCA operator since the original data are first PCA transformed and then fed into the autoencder. Therefore we use inverse PCA to map the reconstructed points back to the gene space for evaluation.

E.2. Generating along geodesics

E.2.1. Datasets: Simulated manifolds

We generate four toy manifolds: ellipsoid, torus, saddle, and hemisphere in $\mathbb{R}^3$. We add Gaussian noise of different scales to the original toy manifolds and rotate the data to higher dimensions using a random rotation matrix. We simulate datasets under $\{0, 0.1, 0.3, 0.5\}$ noise scales and $\{3, 5, 10, 15\}$ dimensions. For each dataset, we randomly select 20 pairs of starting and ending points on the manifold.

We benchmark all methods on the noisy, high-dimensional data, and compute the pairwise geodesics.

E.2.2. Evaluation Criteria

Quantitatively, we evaluate these methods on the MSE criteria: the mean squared error between the predicted geodesic length and ground truth length.

$$\text{Length MSE} = \frac{1}{k} \sum_{i=1}^{k} (\hat{l}_i - l_i)^2, \tag{E.28}$$

where k is the total number of geodesics, $l_i, \hat{l}_i$ are the lengths of the i-th ground truth and predicted geodesics. We obtain the ground truth geodesics analytically if the solution is available or using Dijkstra's algorithm on noiseless data otherwise.

E.3. Geodesics-guided flow matching

E.3.1. Datasets: Randomly sampled populations on toy manifolds

To showcase GAGA's ability on transporting distributions on manifolds, we generate four toy manifolds: ellipsoid, torus, saddle, and hemisphere in $\mathbb{R}^3$. To simulate starting and ending distributions, we first randomly sample two points on the manifold as the starting and ending center and then sample N points near these selected centers. We compute and visualize the flow paths between the two distributions.

F. Additional Experiment Results

F.1. Geometry-aware autoencoder under increasingly noisy data

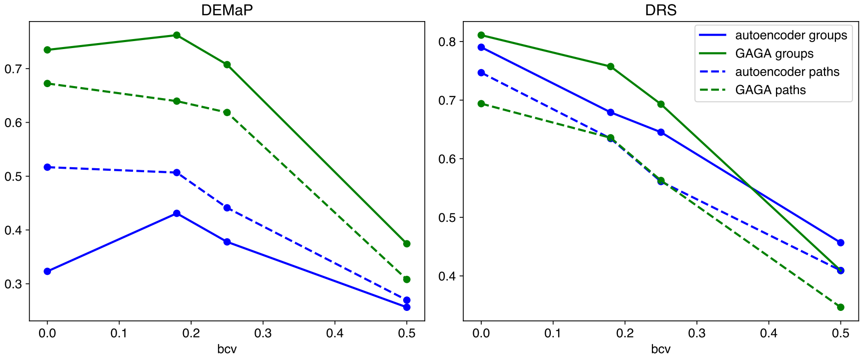

Figure 4. Comparison for GAGA and standard autoencoder on increasingly noisy single-cell datasets.

In Figure 4, we observe that GAGA consistently outperforms standard autoencoder on DEMaP under increasingly noisy sinle-cell data simulated with increasing bcv parameter. Moreover, we can see that GAGA generally rivals the standard autoencoder on DRS, indicating our distance-matching loss does not detract from data reconstruction.

F.2. Visualizing GAGA's latent embeddings

Qualitatively, we visualize the latent embeddings of GAGA on real-world scRNA-seq dataset EB, embryoid body data generated over 27 day time course (Moon et al., 2019). We show that GAGA is able to capture geometric structures in the data, which are essential for biological insights and interpretations. In addition to PHATE, we trained GAGA with two other geodesic distances obtained under different settings of HeatGeo (Huguet et al., 2024). We can see from Figure 5 that GAGA captures both local and global geometric structures such as clusters, branches, and paths. Moreover, Figure 5 shows that GAGA can match closely with the embedding method that it's based on, preserving the latent space of the original dimension reduction method and, at the same time, capable of generalizing to unseen points.

F.3. Visualizing geodesics on toy manifolds

Figure 6 shows the geodesics of different methods on the same set of starting and ending points on multiple toy manifolds. Each row corresponds to one manifold and each column corresponds to one method. From left to right column, the method is 1) ground truth, 2) GAGA, 3) local metric, 4) density regularization. Density refers to geodesics learned with using density regularization.

We can see that GAGA generally outperforms all the other methods except Djikstra's on the saddle datasets. Directly using the local metric performs the worst, lagging far behind all other methods. The inferior performance of the local metric again illustrates the challenges of staying on the manifold while optimizing for the shortest path.

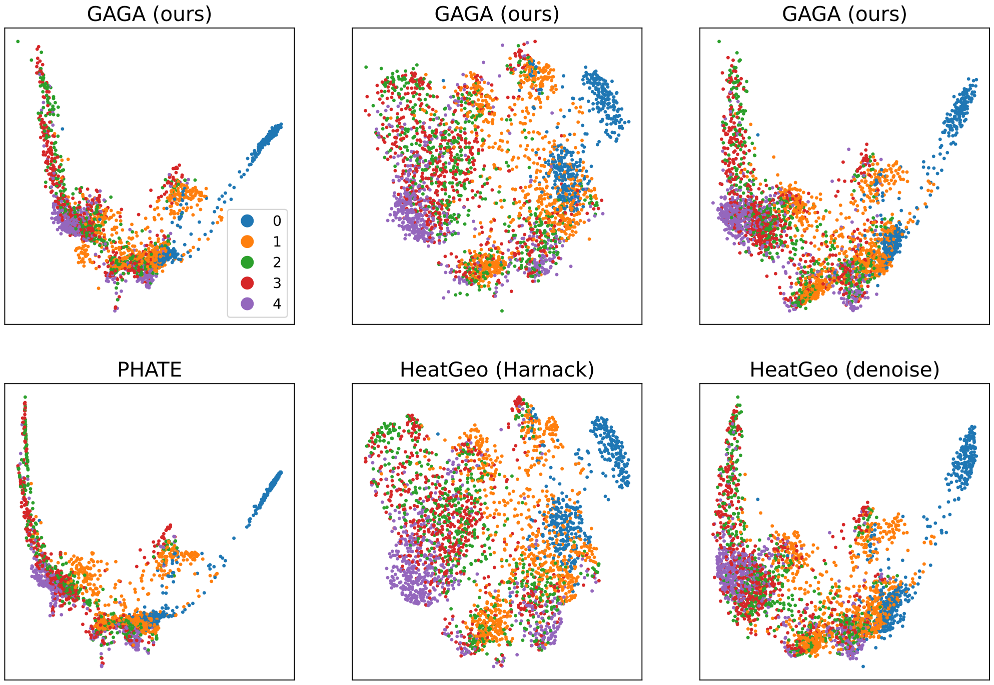

Figure 5. Visualization of the embedding shows GAGA preserves local and global structures.

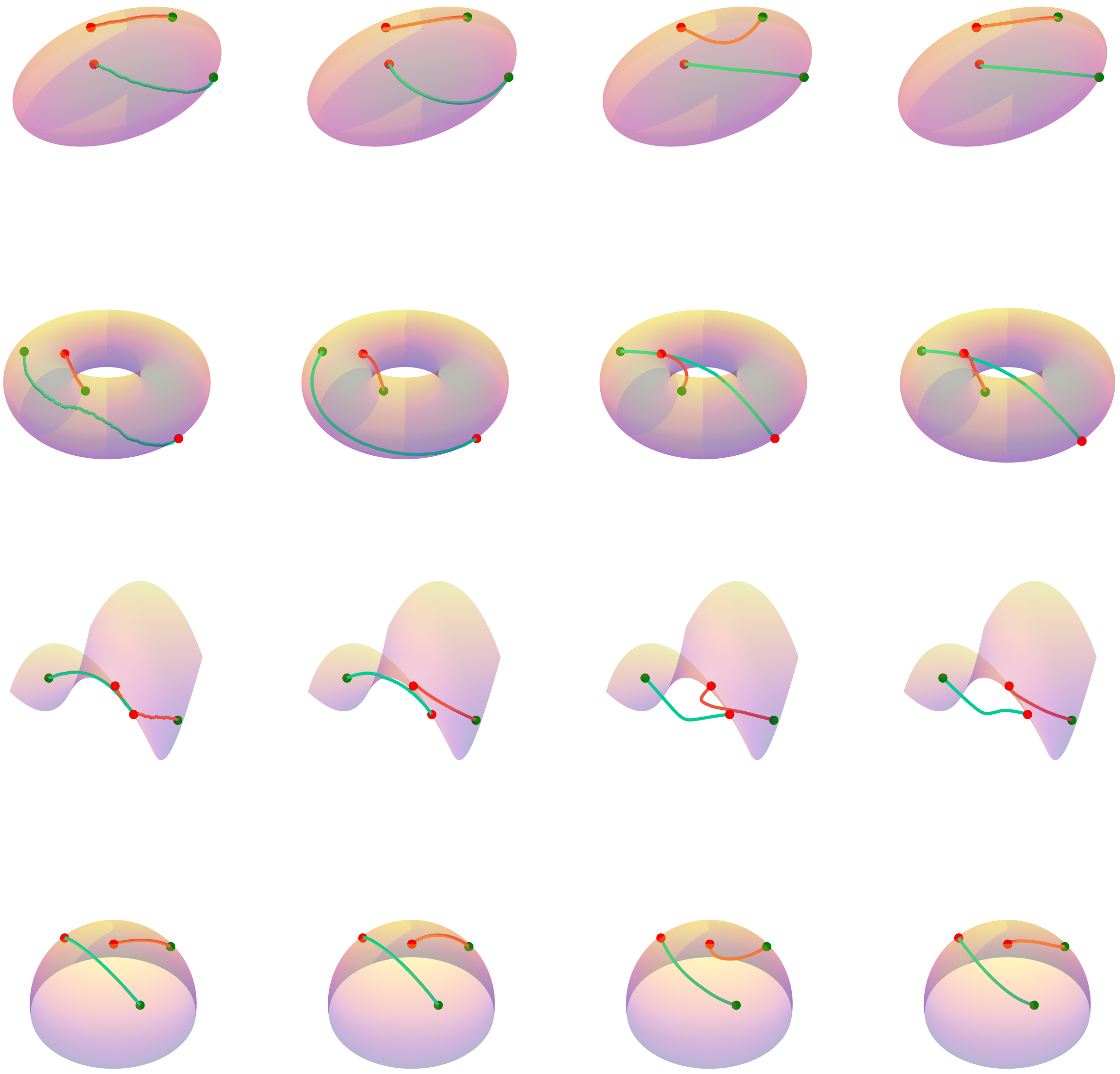

Figure 6. Comparison of geodesics. From left to right columns: 1) ground truth, 2) GAGA, 3) local metric, 4) density regularization.