Wasserstein Flow Matching: Generative Modeling Over Families of Distributions

Doron Haviv [*1,2] **Aram-Alexandre Pooladian** [*3] **Dana Pe'er** [2,4] **Brandon Amos** [5]

Abstract

Generative modeling typically concerns transporting a single source distribution to a target distribution via simple probability flows. However, in fields like computer graphics and single-cell genomics, samples themselves can be viewed as distributions, where standard flow matching ignores their inherent geometry. We propose Wasserstein flow matching (WFM), which lifts flow matching onto families of distributions using the Wasserstein geometry. Notably, WFM is the first algorithm capable of generating distributions in high dimensions, whether represented analytically (as Gaussians) or empirically (as point-clouds). Our theoretical analysis establishes that Wasserstein geodesics constitute proper conditional flows over the space of distributions, making for a valid FM objective. Our algorithm leverages optimal transport theory and the attention mechanism, demonstrating versatility across computational regimes: exploiting closed-form optimal transport paths for Gaussian families, while using entropic estimates on point-clouds for general distributions. WFM successfully generates both 2D & 3D shapes and high-dimensional cellular microenvironments from spatial transcriptomics data. Code is available at Wasserstein Flow Matching.

1. Introduction

Today's abundance of data and scalability of training massive neural networks has made it possible to generate hyper-realistic images on the basis of training examples (OpenAI, 2022), as well as video and audio clips (Vyas et al., 2023; Xing et al., 2023), and, of course, text (Bubeck et al., 2023). All of these are instances of generative modeling: given access to finitely many samples from a distribution, devise a scheme which generates new samples from the same distribution. Generative modeling has also been revolutionary in the biomedical sciences, for drug design (Jumper et al., 2021), and single-cell genomics (Lopez et al., 2018). Nearly all frameworks exploit the notion that datasets (of, say, genomic profiles of cells, images, videos, or corpora of text documents) are instantiations of probability measures, and the task is to transform a point sampled from random noise to generate a data point that obeys the distribution of interest.

Among the zoo of available generative models, one approach noted for its flexibility and simplicity is Flow Matching (FM) (Albergo & Vanden-Eijnden, 2022; Lipman et al., 2022; Liu et al., 2022). For a fixed target probability measure, FM learns an implicitly defined vector field that can transform a source measure (e.g., the standard Gaussian) to the target. Unlike discrete time and probabilistic generative models (such as Diffusion Models by Song et al. (2020)), FM learns a deterministic, continuous normalizing flow by regressing onto a simple conditional probability flow. This approach, while originally designed for Euclidean domains, can be readily adopted to Riemannian geometries (Chen & Lipman, 2023). Riemannian flow matching (RFM) is widely used for generating samples over geometries such as spheres, tori, translation/rotation groups, simplices, triangular meshes, mazes, and molecular positions and structures.

However, modern data can exhibit a richer structure in which the samples themselves are distributions. In computational graphics, 3D models are shape distributions observed through point-clouds. Likewise, recent developments in single-cell genomics analysis have demonstrated that gene-expression profiles from groups of cells aggregated via their mean and covariance can capture cellular microenvironments or highlight fine-grain clusters (Haviv et al., 2024b; Persad et al., 2023). For both general distributions and Gaussian settings, it is natural to search for a unified generative model that respects the underlying geometry of the data.

Contributions. To this end, we introduce *Wasserstein Flow Matching* (WFM), a principled extension of the FM framework lifted to the space of probability distributions. As illustrated in Figure 1, a single point in our source and target datasets is itself a distribution. These distributions are either represented analytically as Gaussians or realized

[*]Equal contribution [1]Weill Cornell [2]Memorial Sloan–Kettering Cancer Center [3]Center for Data Science, New York University [4]Howard Hughes Medical Institute [5]Meta AI. Correspondence to: Doron Haviv <havivd@mskcc.org>, Aram-Alexandre Pooladian <ap6599@nyu.edu>, Brandon Amos <bda@meta.com>.

Proceedings of the 42^{nd} International Conference on Machine Learning, Vancouver, Canada. PMLR 267, 2025.

through point-cloud samples. Our aim is to learn vector fields acting on the space of probability distributions and match the optimal transport map, which is the geodesic in Wasserstein space. WFM is an instantiation of Riemannian FM (Chen & Lipman, 2023), where we train a neural model to learn a continuous normalizing flow (CNF) between distributions over distributions.

We demonstrate the effectiveness of our approach for generative modeling between distributions over Gaussian distributions and distributions over point-clouds, which are realization of general distributions. The former task is motivated by recent directions in single-cell and spatial transcriptomics (Haviv et al., 2024b; Persad et al., 2023), where we consider matching problems over the *Bures–Wasserstein space* (BW), the Gaussian submanifold of the Wasserstein space. In this case, we show that WFM can be further modified, resulting in the Bures–Wasserstein FM (BW-FM) algorithm. We validate BW-FM on a variety of Gaussian-based datasets, where we observe that samples generated by our algorithm are significantly more robust than naïve approaches which do not fully exploit the underlying geometry of the data. As an application, we present a generative model for cell states and niches from single-cell genomics data.

Method	Data type	Source	Target
FM over $\mathbb{R}^d$	$x \in \mathbb{R}^d$	$x \sim p_0$	$y \sim p_1$
FM over $\mathcal{M}$	$x \in \mathcal{M}$	$x \sim \mathfrak{p}_0$	$y \sim \mathfrak{p}_1$
FM over Δ_d	$\mu \in \mathcal{P}(\Delta_d)$	$\mu \sim \mathfrak{p}_0$	$\nu \sim \mathfrak{p}_1$
Wasserstein FM	$\mu \in \mathcal{P}(\mathbb{R}^d)$	$\mu \sim \mathfrak{p}_0$	$\nu \sim \mathfrak{p}_1$
$\to$ Gaussians	$\mathcal{N}(m, \Sigma)$	$\mathcal{N}(m_\mu, \Sigma_\mu) \sim \mathfrak{p}_0$	$\mathcal{N}(m_\nu, \Sigma_\nu) \sim \mathfrak{p}_1$
$\to$ Point-Clouds	$\frac{1}{n}\sum_i \delta_{x_i}$	$\frac{1}{m}\sum_i \delta_{x_i} \sim \mathfrak{p}_0$	$\frac{1}{n}\sum_j \delta_{y_j} \sim \mathfrak{p}_1$

Table 1. Contrasting FM methods over $\mathbb{R}^d$, general manifolds $\mathcal{M}$, categorical and Dirichlet distributions on the d-simplex Δ_d, and finally, our approach, FM problems defined over $\mathcal{P}(\mathbb{R}^d)$.

Generation of general distributions is made possible by two distinct, yet crucial, algorithmic primitives: (1) incorporating transformers in our neural network architecture (Vaswani, 2017; Lee et al., 2019), and (2) recent algorithmic advances in entropic optimal transport (Pooladian & Niles-Weed, 2021). Our WFM algorithm performs generative modeling in the Wasserstein space, where geodesics are given by pushforwards of optimal transport (OT) maps; see Section 2.3 for more information. Since these OT maps lack closed-form solutions for general distributions, we represent distributions as point-clouds and estimate the maps using entropic optimal transport. The permutation equivariance of attention makes transformers a natural basis for our model, inherently capturing the equivariance of Wasserstein geometry while maintaining scalability in high dimensions.

For datasets of distributions in 3D, WFM's performance matches existing generative models. However, due to their reliance on voxelization, current approaches cannot scale to high-dimensional distributions and fail when distributions

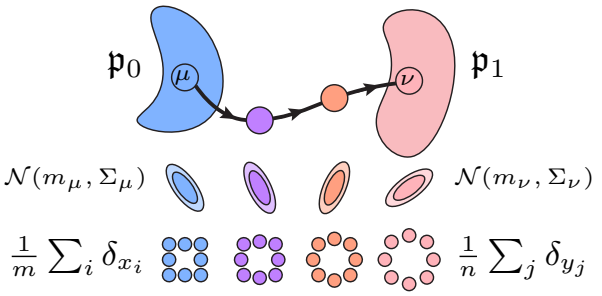

Figure 1. WFM learns flows between distributions over distributions, where distributions are either analytically represented (Gaussians) or empirically observed through point-clouds.

are realized as point-clouds with variable sizes. Conversely, WFM succeeds in these challenging settings, enabling generative modeling in new domains like synthesizing tissue microenvironments from spatial genomics data. Modeling tissue biology in this generative manner could enhance our understanding of how environment is associated with cell state. In the context of many diseases, most notably cancer and its immune microenvironment, these insights are critical for developing novel therapeutics (Binnewies et al., 2018).

2. Background and related work

We let $\mathcal{P}_2(\mathbb{R}^d)$ denote probability distributions over $\mathbb{R}^d$ with finite second moment, and write $\mathcal{P}_{2,\text{ac}}(\mathbb{R}^d)$ for those with densities. For a probability measure μ and function f, we write $|f|^2_{L^2(\mu)}$ for the squared $L^2(\mu)$ norm. Let $\mathcal{M}$ be a Riemannian manifold with tangent space $\mathcal{T}_x\mathcal{M}$ at x. For $x_0 \in \mathcal{M}$ with velocity $v \in \mathcal{T}_{x_0}\mathcal{M}$, the exponential map $\exp_{x_0}(v)$ gives the terminal location, while the logarithmic map $\log_{x_0}(x_1)$ gives the initial velocity for the geodesic from x_0 to x_1. The set of symmetric (resp. positive definite) matrices over $\mathbb{R}^d$ are denoted by $\mathbb{S}^d$ (resp. $\mathbb{S}^d_{++}$).

2.1. Riemannian flow matching

We first briefly discuss the Riemannian flow matching (RFM) framework of Chen & Lipman (2023). Let $\mathfrak{p}_0$ be the source distribution and $\mathfrak{p}_1$ be the target distribution over a Riemannian manifold $\mathcal{M}$, and let $(\gamma_t)_{t\in[0,1]}$ be a curve of probability measures satisfying $\gamma_0 = \mathfrak{p}_0$ and $\gamma_1 = \mathfrak{p}_1$. Letting $(w_t)_{t\in[0,1]}$ denote a family of vector fields, we say that the pair $(\gamma_t, w_t)_{t\in[0,1]}$ satisfy the *continuity equation* with respect to the metric g, abbreviated to $(\gamma_t, w_t) \in \mathfrak{C}_g$ if

$$\partial_t \gamma_t + \nabla_g \cdot (\gamma_t w_t) = 0\,, \tag{1}$$

where $\nabla_g \cdot$ is the Riemannian divergence operator.

The goal of RFM is to regress a parameterized vector field (e.g., a neural network), written $f_\theta(x,t) \in \mathcal{T}_x\mathcal{M}$ for $t \in$

$[0, 1]$, onto the family w_t by minimizing

$$\min_\theta \int_0^1 \int \|f_\theta(z_t, t) - w_t(z_t)\|^2_{g(z_t)} \, \mathrm{d}\gamma_t(z_t) \, \mathrm{d}t \,,$$

assuming access to a pair $(\gamma_t, w_t)_{t\in[0,1]}$ that satisfies (1). This is not possible in many scenarios. Borrowing insights from recent work (e.g., Albergo & Vanden-Eijnden (2022); Lipman et al. (2022); Liu et al. (2022)), the authors construct a simple vector field that satisfies the continuity equation, resulting in the tractable objective

$$\min_\theta \int_0^1 \iint \|f_\theta(x_t, t) - \dot{x}_t\|^2_{g(x_t)} \, \mathrm{d}\mathfrak{p}_0(x) \, \mathrm{d}\mathfrak{p}_1(y) \, \mathrm{d}t \,, \quad (2)$$

where, for example, $x_t = \exp_x((1-t)\log_x(y)) \in \mathcal{M}$, and $\dot{x}_t \in \mathcal{T}_{x_t}\mathcal{M}$. For complete discussions and proofs, see Chen & Lipman (2023, Section 3.1). Once f_θ is appropriately fit using (2), we can generate new samples from $\mathfrak{p}_1$: start by sampling $X_0 \sim \mathfrak{p}_0$, then follow $\dot{X}_t = f_\theta(X_t, t)$ numerically by discretizing the dynamics given by the exponential map, resulting in $X_1 \sim \mathfrak{p}_1$. We emphasize that the dynamics are only simulated at inference time and not when training f_θ, commonly known as a *simulation-free* training paradigm.

2.2. Related work

Generative models for shapes. Paralleling the progress in generative models for natural images, the field of shape generation is rapidly expanding. Many different models have been used from this task, namely generative-adversarial-nets (Achlioptas et al., 2018), variational autoencoders (Gadelha et al., 2018), normalizing flows (Yang et al., 2019; Kim et al., 2020; Klokov et al., 2020), diffusion (Zhou et al., 2021; Cai et al., 2020) and even euclidean FM (Wu et al., 2023a). Thus far, these approaches are limited to uniform-sized samplings of shape distributions in 2D & 3D, and fail on datasets where realizations are variably sized.

Generative models over families of distributions. Our work is not the first to instantiate Riemannian FM with a manifold of probability measures. Two notable works are Fisher FM (Davis et al., 2024) and Categorical FM (Cheng et al., 2024), which consider the FM algorithm with respect to the Fisher–Rao geometry (Amari, 2016; Nielsen, 2020) over the d-dimensional simplex Δ_d. The work of Stark et al. (2024) is similar in spirit, where they focus on the Dirichlet distribution for generation of discrete data. Another related work is that of Atanackovic et al. (2024), called Meta FM. Their approach requires pairs of distributions which are already coupled, with the goal of solving FM between a distribution over pairs. In contrast, we emphasize that our proposed Wasserstein FM applies between two separate *uncoupled* distributions over distributions.

Generative models for single-cell genomics. Deep learning based generative models have transformed single-cell genomics through various approaches. Variational autoencoders (Lopez et al., 2018; Gayoso et al., 2022) have successfully addressed technical artifacts, integrating multimodal data and imputing missing features in scRNA-seq data. More recently, Transformer-based foundation models have been noted for their ability to integrate large atlases of data (Cui et al., 2024; Theodoris et al., 2023). Flow Matching has also emerged as a promising direction (Klein et al., 2024; Eyring et al., 2024) for learning both balanced and unbalanced OT maps between cell populations. While these prior FM applications focus on cell-to-cell mappings, our work introduces a new paradigm: generating entire distributions representing cellular populations. This is particularly relevant for spatial genomics, where cellular microenvironments are naturally represented as distributions. WFM enables synthesis of whole cellular neighborhoods, a novel approach in generative modeling for the single-cell field.

2.3. Wasserstein geometry

The (squared) 2-*Wasserstein distance* between two probability measures $\mu, \nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ is given by the optimization problem over vector-valued maps $T : \mathbb{R}^d \to \mathbb{R}^d$

$$W_2^2(\mu, \nu) \coloneqq \min_{T:T_\sharp\mu=\nu} \|\mathrm{id} - T\|^2_{L^2(\mu)} \,, \quad (3)$$

where the pushforward constraint, written $T_\sharp\mu = \nu$, means that for $X \sim \mu$, $T(X) \sim \nu$. The minimizer to (3) is called the *optimal transport (OT) map*, denoted $T_\star^{\mu\to\nu}$ (we abbreviate this to $T_\star$ when clear). The existence and uniqueness of the optimal transport map under the stated regularity conditions is due to Brenier (1991).

The *Wasserstein space* is the space of probability densities with finite second moment endowed with the Wasserstein distance; this space is known to be a metric space (Villani, 2009). Following the celebrated work of (Otto, 2001), the Wasserstein space can be formally (meaning, non-rigorously) viewed as a Riemannian manifold, whose properties we now describe in brief; see e.g., (Ambrosio et al., 2008) for a rigorous treatment. Following the definition by Ambrosio et al. (2008, Theorem 8.5.1), the tangent space at a point $\mu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$ consists of all possible vectors that emanate from μ, written as

$$\mathcal{T}_\mu\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d) \coloneqq \overline{\{\lambda(T_\star^{\mu\to\nu} - \mathrm{id}) : \lambda > 0, \nu \in \mathcal{P}_2(\mathbb{R}^d)\}}^{L^2(\mu)}$$

where the overline denotes the closure of the set (i.e., the set and its limit points) in $L^2(\mu)$, and the norm on the tangent space is $L^2(\mu)$. The exponential and logarithmic maps read

$$v \mapsto \exp_\mu(v) \coloneqq (\mathrm{id} + v)_\sharp\mu \,, \; \nu \mapsto \log_\mu(\nu) \coloneqq T_\star^{\mu\to\nu} - \mathrm{id} \,,$$

where id is the identity map. Consequently, the (constant-speed) geodesic, or *McCann interpolation*, between two

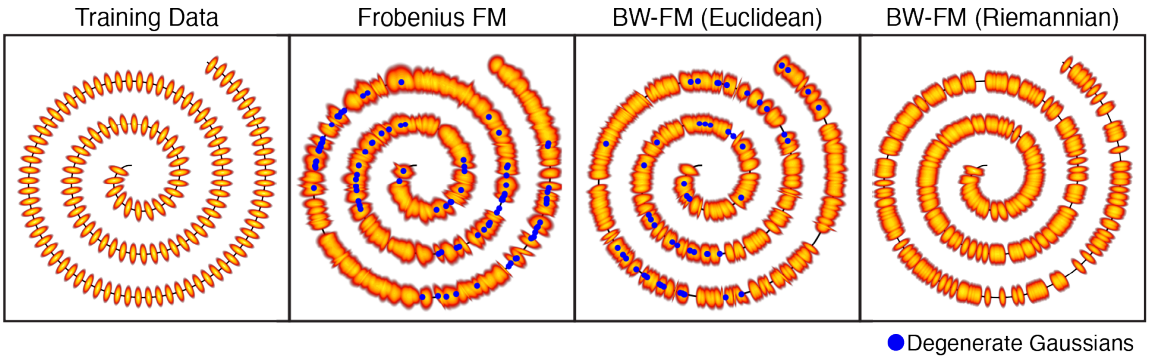

Figure 2. In the presence of sufficiently many samples, all methods generate Gaussians along the whole spiral, and our Riemannian BW-FM algorithm produces the most consistent samples. Other methods produce Gaussians with *degenerate* covariance, as they do not model geometry of the data. When there are only few examples, BW-FM accurately reconstructs the training data; see Figure S1.

measures μ and ν is given by the curve $(\mu_t)_{t\in[0,1]}$ where

$$\begin{aligned}\mu_t &:= (T_t^{\mu\to\nu})_\sharp\mu := ((1-t)\mathrm{id} + tT_\star^{\mu\to\nu})_\sharp\mu \\ &\equiv \exp_\mu(t\log_\mu(\nu))\,,\end{aligned} \tag{4}$$

where the last expression writes the pushforward in terms of the exponential and logarithmic maps. Equivalently, at the level of the random variables, one can write $X_t = (1-t)X_0 + tT_\star^{\mu\to\nu}(X_0)$, where $X_0 \sim \mu$ and $X_t \sim \mu_t$ for any $t \in [0,1]$. Combined with $(v_t)_{t\in[0,1]}$ a suitable family of vector fields, the McCann interpolation satisfies the continuity equation (1) over $\mathbb{R}^d$, re-written as

$$\partial_t\mu_t + \nabla\cdot(\mu_t v_t) = 0\,, \quad \text{s.t.} \quad \mu_0 = \mu\,, \mu_1 = \nu\,, \tag{5}$$

where the divergence operator is the usual Euclidean one over $\mathbb{R}^d$, thus we write $(\mu_t, v_t) \in \mathfrak{C}$. The link between the constant speed geodesics and the 2-Wasserstein distance can be viewed from the celebrated Benamou–Brenier formulation of optimal transport (Benamou & Brenier, 2000):

$$W_2^2(\mu,\nu) = \inf_{(\mu_t,v_t)\in\mathfrak{C}} \int_0^1 \|v_t\|^2_{L^2(\mu_t)}\,\mathrm{d}t\,. \tag{6}$$

The optimal curve of measures is given by the constant-speed geodesics described above, and the optimal velocity field is given by

$$v_t = (T_\star^{\mu\to\nu} - \mathrm{id})\circ(T_t^{\mu\to\nu})^{-1}\,. \tag{7}$$

The vector field (7) should be interpreted as the time-derivative of the McCann interpolation, with $X_0 \sim \mu$

$$\begin{aligned}\dot X_t &= (T_\star^{\mu\to\nu} - \mathrm{id})(X_0) \\ &= (T_\star^{\mu\to\nu} - \mathrm{id})\circ(T_t^{\mu\to\nu})^{-1}(X_t)\,.\end{aligned}$$

2.3.1. Bures–Wasserstein (BW) space

A known special case of the Wasserstein space is the *Bures–Wasserstein* space, which consists of the submanifold of non-degenerate Gaussians parameterized by means and covariances $\{(m,\Sigma) : m \in \mathbb{R}^d, \Sigma \in \mathbb{S}^d_{++}\}$, endowed with the Wasserstein metric. We provide a brief exposition on the geometry of the Bures–Wasserstein space and refer to Lambert et al. (2022) for detailed calculations and explanations, as we follow their notation conventions.

The OT map between $\mathcal{N}(m_\mu,\Sigma_\mu)$ and $\mathcal{N}(m_\nu,\Sigma_\nu)$ has a closed-form expression (Gelbrich, 1990):

$$\begin{aligned}T_\star(x) &:= m_\nu + C^{\mu\to\nu}(x - m_\mu) \\ &:= m_\nu + \Sigma_\mu^{-\frac12}(\Sigma_\mu^{\frac12}\Sigma_\nu\Sigma_\mu^{\frac12})^{\frac12}\Sigma_\mu^{-\frac12}(x-m_\mu)\,.\end{aligned}$$

As this map is affine, it is clear that the McCann interpolation between two Gaussians is always Gaussian. More generally, we have the succinct representation of the tangent space at a point in the Bures–Wasserstein space

$$\mathcal{T}_\mu\mathrm{BW}(\mathbb{R}^d) := \{a + S(\mathrm{id} - m_\mu) \,:\, a\in\mathbb{R}^d, S\in\mathbb{S}^d\}\,,$$

and the tangent space norm at μ can be written as:

$$\|(a,S)\|^2_{\mathrm{BW}(\mu)} := \|a\|^2 + \mathrm{Tr}(S^2\Sigma_\mu)$$

With the above, it is straightforward to compute the McCann interpolation $\mu_t = (T_t)_\sharp\mu = \mathcal{N}(m_t,\Sigma_t)$, with

$$\begin{aligned}m_t &:= (1-t)m_\mu + tm_\nu\,, \\ \Sigma_t &:= T_tAT_t,\ T_t = (1-t)I + tC^{A\to B}\end{aligned} \tag{8}$$

We can relate the Euclidean and Riemannian time-derivatives of Σ_t through the following manipulation:

$$\dot\Sigma_t^{\mathrm{E}} = \dot T_tAT_t + T_tA\dot T_t = \dot\Sigma_t^{\mathrm{BW}}\Sigma_t + \Sigma_t\dot\Sigma_t^{\mathrm{BW}}\,. \tag{9}$$

To this end, we can draw parallels to (7) by writing

$$\begin{aligned}\dot m_t &= m_\nu - m_\mu\,, \\ \dot\Sigma_t^{\mathrm{BW}} &= (C^{A\to B} - I)((1-t)I + tC^{A\to B})^{-1}\,.\end{aligned} \tag{10}$$

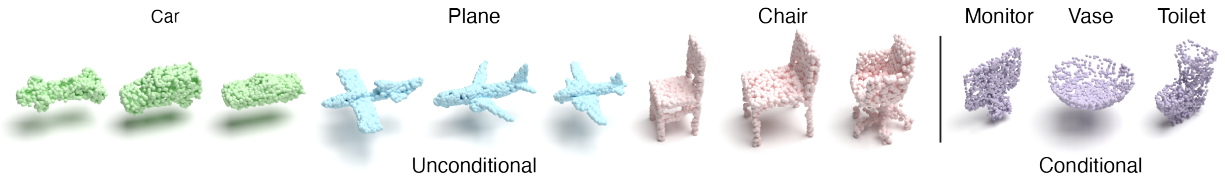

Figure 3. Left. Synthesized samples from WFM trained on the *cars*, *planes* or *chairs* datasets. *Right.* Examples generated conditionally from the *same initial noise* via a WFM model trained on the complete 40-class ModelNet dataset.

3. Flow matching over the Wasserstein space

Let $\mathfrak{p}_0$ and $\mathfrak{p}_1$ denote probability measures over the Wasserstein space. Our goal is to learn a vector field that transports the family of measures $\mathfrak{p}_0$ to the family $\mathfrak{p}_1$. WFM learns to map source to target by regressing onto Wasserstein geodesics between samples $\mu \sim \mathfrak{p}_0$ and $\nu \sim \mathfrak{p}_1$.[1] To accomplish this, we pass in the McCann interpolation μ_t and optimal velocity field v_t in the Riemannian FM objective (2), resulting in our Wasserstein FM (WFM) objective:

$$\min_\theta \int_0^1 \iint \|f_\theta^{\text{geo}}(\mu_t, t) - v_t\|^2_{L^2(\mu_t)} \, \mathrm{d}\mathfrak{p}_0 \, \mathrm{d}\mathfrak{p}_1 \, \mathrm{d}t \,. \quad (11)$$

See Appendix B for a derivation of this training objective. It remains to ensure that this objective describes valid conditional probability flows between families of distributions.

First and foremost, we require two levels of regularity on the distributions of distributions $\mathfrak{p}_0$ and $\mathfrak{p}_1$:

1. **Outer continuity:** The source measure $\mathfrak{p}_0$ satisfies smoothness conditions (as in Appendix A of Chen et al. (2024)) ensuring the conditional paths $\mathfrak{p}_t(\cdot|\nu)$ can be marginalized over $\mathfrak{p}_1$ to form a valid flow
2. **Inner continuity:** For any $\mu \sim \mathfrak{p}_0$ and $\nu \sim \mathfrak{p}_1$, there exists an optimal transport map between μ and ν making the conditional flows well-defined.

The condition of "inner continuity" is fairly mild, as this is ensured for any distribution $\mu \sim \mathfrak{p}_0$ with density. For Gaussian distributions, inner continuity holds naturally. For general distributions, we assume continuity but work with point-clouds as empirical realizations to approximate OT maps with statistical guarantees (see Appendix A.3) and computational efficiency (Flamary et al., 2021; Cuturi et al., 2022). There exists a slight gap between our theoretical results which consider classic OT and our practical implementation which uses entropy-regularized OT approximations. This aligns with other geometric methods (see Algorithm 1 by Chen & Lipman (2023)) that rely on numerical approximations when exact solutions are not tractable. The "outer continuity" condition is purely technical, and it serves the same role as in prior work. Our training algorithm is described in Algorithm 1, and Appendix E contains precise details on our neural network design.

[1] Crucially, we are *not* learning the OT map between $\mathfrak{p}_0$ and $\mathfrak{p}_1$.

Finally, we prove that regressing onto vector fields results in valid conditional probability flows between distributions of distributions; see Appendix B.1 for a proof.

Proposition 3.1. *Conditional probability paths $\mathfrak{p}_t(\cdot|\nu)$ generated by conditional vector fields of the form v_t (recall (7)) satisfy $\mathfrak{p}_1(\cdot|\nu) = \nu$.*

To prove this result, we following a general recipe by Chen & Lipman (2023) with calculations specific to the Wasserstein space. First, we prove that the Wasserstein distance is a *pre-metric* (see Lemma B.1). Then, we prove that the conditional vector fields v_t are defined as a function of this pre-metric (see Lemma B.2). With these calculations, the proof of Proposition 3.1 then follows immediately from Theorem 3.1 of Chen & Lipman (2023).

3.1. WFM over the Bures–Wasserstein space

First suppose $\mathfrak{p}_0$ and $\mathfrak{p}_1$ are distributions over Gaussians, meaning that a batch of samples drawn from $\mathfrak{p}_0$ and $\mathfrak{p}_1$ consists of a batch of mean-covariance pairs. Here, the dynamics are straightforward: the interpolant is the McCann interpolation, and the velocity field over the Bures–Wasserstein manifold is also known (see (8) & (10), respectively). Since μ_t is (m_t, Σ_t), the neural network is parameterized as $f_\theta^{\text{BW}} : \mathbb{R}^d \times \mathbb{S}^d_{++} \to \mathbb{R}^d \times \mathbb{S}^d$, and the norm on the tangent space simplifies the computations considerably. Our final training objective becomes

$$\int_0^1 \iint \|f_\theta^{\text{BW}}((m_t, \Sigma_t), t) - (\dot{m}_t, \dot{\Sigma}_t^{\text{BW}})\|^2_{\text{BW}} \, \mathrm{d}\mathfrak{p}_0 \, \mathrm{d}\mathfrak{p}_1 \, \mathrm{d}t \,.$$

3.2. WFM over distributions of general distributions

In the case of general distributions, we lose closed-form interpolations. However, we can hope to proceed so long as we have an *approximation* of the optimal transport map between them, written $\hat{T}$, based on empirical samples (point-clouds) drawn from the distributions. There are many works on the approximation of these maps on the basis of samples; see Hütter & Rigollet (2021); Divol et al. (2022); Manole et al. (2021); Pooladian & Niles-Weed (2021).

While existing point-cloud generative models work with uniform-sized samples from high-resolution CAD models, many datasets provide inherently variable-sized samples. Though we assume these samples come from underlying continuous distributions, we only observe finite, discrete

Algorithm 1 Wasserstein FM Training

Data: base $\mathfrak{p}_0 \in \mathcal{P}(\mathcal{P}(\mathbb{R}^d))$, target $\mathfrak{p}_1 \in \mathcal{P}(\mathcal{P}(\mathbb{R}^d))$, geo $\in \{\text{BW}, \text{General}\}$
Init: Parameters θ of f_θ^{geo}
repeat
  Sample time $t \sim \mathcal{U}(0,1)$
  Sample source measure $\mu \sim \mathfrak{p}_0$
  Sample target measure $\nu \sim \mathfrak{p}_1$
  if geo is BW **then**
    $\mu_t \leftarrow (m_t, \Sigma_t)$ via equation 8
    $v_t \leftarrow (\dot{m}_t, \dot{\Sigma}_t^{\text{BW}})$ via equation 9
  else
    $\mu_t \leftarrow$ Approximate via equation 4
    $v_t \leftarrow$ Approximate via equation 7
  end if
  $\ell(\theta) \leftarrow \|f_\theta^{\text{geo}}(\mu_t, t) - v_t\|^2_{L^2(\mu_t)}$
  $\theta \leftarrow \text{optimizer\_step}(\theta, \ell(\theta), \nabla_\theta \ell(\theta))$
until Convergence

realizations. This natural variation in sample sizes necessitates our development of WFM to handle point-clouds of non-uniform size, making WFM applicable across both high-resolution shape modeling and naturally discrete datasets.

We approximate optimal transport maps using entropic OT (Cuturi, 2013), which is computationally efficient on GPUs via Sinkhorn's algorithm (Sinkhorn, 1964). Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ represent point-clouds described by data from $\mu \sim \mathfrak{p}_0$ and $\nu \sim \mathfrak{p}_1$ respectively, and let $\hat{T}^{\mu\to\nu}$ denote the *entropic* approximation of the optimal transport map (see Appendix A for full details). The objective (11) can be approximated by

$$\int_0^1 \iint \|f_\theta^{\text{PC}}(\hat{\boldsymbol{X}}_t, t) - ((\hat{T}^{\mu\to\nu}(\boldsymbol{X}) - \boldsymbol{X}))\|_2^2 \, \mathrm{d}\mathfrak{p}_0 \, \mathrm{d}\mathfrak{p}_1 \, \mathrm{d}t \,,$$

where $\hat{\boldsymbol{X}}_t = (1-t)\boldsymbol{X} + t\hat{T}^{\mu\to\nu}(\boldsymbol{X})$ represents a discretized McCann interpolation. We parameterize f_θ^{PC} with a transformer, leveraging the natural alignment between the permutation equivariance of both OT maps and self-attention. Unlike point-cloud models based on voxelization, transformers maintain efficiency in high dimensions.

3.3. Generation

Once f_θ^{geo} is trained, we can generate new samples as in Riemannian FM. For the Bures–Wasserstein space, we appeal to the closed-form exponential and logarithmic maps; see Algorithm 2. We emphasize that the appropriate Riemannian updates are crucial to obtain non-degenerate final samples. For general distributions, we perform a standard Euler discretization of the learned flow on their point-cloud realizations; see Algorithm 3.

Algorithm 2 $\text{BW}(\mathbb{R}^d)$ Generation

Data: Trained f_θ^{BW}, step size $h = 1/N$
Init: $\mathcal{N}(m_0, \Sigma_0) \sim \mathfrak{p}_0$
for $k = 0, \ldots, N-1$ **do**
  $(s_k, S_k) \leftarrow f_\theta^{\text{BW}}((m_{kh}, \Sigma_{kh}), kh)$
  $m_{(k+1)h} \leftarrow m_k + hs_k$, $U_k \leftarrow (I + hS_k)$
  $\Sigma_{(k+1)h} \leftarrow U_k \Sigma_{kh} U_k$
end for
Return: $\mathcal{N}(m_{Nh}, \Sigma_{Nh})$

Algorithm 3 General Distribution Generation

Data: Trained f_θ^{PC}, step size $h = 1/N$
Init: $\mu_0 \sim \mathfrak{p}_0$, $\hat{\mathbf{X}}_0 = \{X_1, \ldots, X_n\} \sim \mu_0$
for $k = 0, \ldots, N-1$ **do**
  $\hat{\mathbf{X}}_{(k+1)h} \leftarrow \hat{\mathbf{X}}_{kh} + hf_\theta^{\text{PC}}(\hat{\mathbf{X}}_{kh}, kh)$
end for
Return: $\hat{\mathbf{X}}_{Nh}$

4. Results

4.1. Families of Gaussians

We first demonstrate our flow matching framework between measures of Gaussians on synthetic and real datasets. For comparable baselines in each scenario, we construct two simpler flow matching approaches for Gaussian generation: (1) *Frobenius FM*, which concatenates the mean and covariances, and trains on $\dot{\Sigma}_t^{\text{E}}$ (equation (9)) with respect to the squared-Frobenius norm, and (2) BW-FM (Euclidean), which tries to match $\dot{\Sigma}_t^{\text{E}}$ but still under the BW geometry.

	BW-FM (R)	BW-FM (E)	Frobenius FM
Spiral - 16 (2D)	$\mathbf{2.98 \cdot 10^{-4}}$	$4.00 \cdot 10^{-4}$	$1.03 \cdot 10^{-3}$
Spirals - 128 (2D)	$\mathbf{1.28 \cdot 10^{-3}}$	$1.70 \cdot 10^{-3}$	$2.69 \cdot 10^{-3}$
Two Moons (2D)	$\mathbf{1.84 \cdot 10^{-4}}$	$8.96 \cdot 10^{-4}$	$1.30 \cdot 10^{-3}$
Sphere (3D)	$\mathbf{6.65 \cdot 10^{-4}}$	$2.14 \cdot 10^{-3}$	$2.25 \cdot 10^{-3}$
Cities (2D)	$\mathbf{1.88 \cdot 10^{-4}}$	$7.26 \cdot 10^{-3}$	$1.75 \cdot 10^{-3}$
ECG (15D)	$\mathbf{9.24 \cdot 10^{-2}}$	$3.26 \cdot 10^{-1}$	$3.98 \cdot 10^{-1}$
seqFISH (16D)	$\mathbf{3.11 \cdot 10^{-1}}$	$5.31 \cdot 10^{-1}$	$6.82 \cdot 10^{-1}$
scRNA-seq (32D)	**1.31**	2.74	3.21

Table 2. Average min. W_2^2 distance between each generated Gaussian and the reference datasets. Despite identical training schemes, BW-FM (R) outperforms other approaches on both synthetic and real data across several dimensions.

In all cases, we assess the quality of the learned flows by computing the the minimum distance between each generated Gaussian and the dataset using the (squared) 2-Wasserstein distance; see Table 2. Notably, the flows generated by Frobenius FM and BW-FM (Euclidean) do not strictly adhere to the geometry of the Bures-Wasserstein manifold, requiring synthesized covariance matrices to be projected onto the space of PSD matrices. In contrast, the Riemannian BW-FM algorithm consistently produces valid and accurate results across all dimensions and datasets.

4.1.1. Toy Datasets

As a first test, we design a dataset of Gaussians centered on a spiral; See Figures 2 & S1. When there were only few samples, only Riemannian BW-FM reconstructed the data, despite other benchmark methods following identical training regimes. On the complete 128-sample dataset, generalizes beyond the training data and synthesizes novel Gaussians lying on the spiral. BW-FM shares this generalization feature with standard FM, and is able to learn the structure underlying the measure from the training data.

4.1.2. Single-Cell Genomics

Spatial transcriptomics are a set of techniques which build on single-cell genomics and preserve physical information of cells' location in tissues and assay their gene expression. Haviv et al. (2024b) demonstrated that a cell's microenvironment can be effectively characterized using the mean and covariance of the surrounding cells gene expression. This representation captures key features of cellular neighborhoods and transforms spatial transcriptomics datasets into a measure within the Bures–Wasserstein space, highlighting the value of generative modeling in this context.

During embryogenesis, specific regions within the primitive gut tube differentiate into organs such as the liver or lungs based on interactions between the gut and surrounding mesenchyme (Nowotschin et al., 2019). Applied on environments of gut-tube cells from a seqFISH dataset of mouse embryogenesis (Lohoff et al., 2022), BW-FM synthesized Gaussian cellular niches conditioned on organ labels, thus demanding an understanding of the interplay between spatial context and phenotype. Despite the intricate nature of the gastrulation process, compounded by the dataset's dimensionality, BW-FM can accurately generate organ-specific niches; see Figure S2.

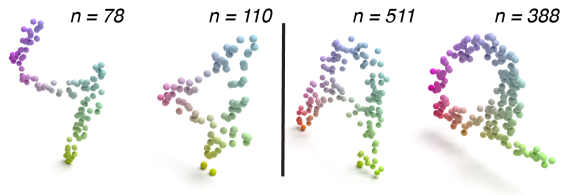

Figure 4. Generated distributions from MNIST and Letters datasets using WFM. Each distribution is realized as a point-cloud with naturally varying sample sizes (n), demonstrating WFM's unique ability to learn from and generate distributions without requiring fixed-size realizations.

Another common instance of BW manifolds arising in single-cell genomics is through aggregating cells into common states. These clusters can be summarized by their mean gene expression and its covariance. On a scRNA-seq atlas elucidating human immune response to COVID (Stephenson et al., 2021), we combine cells into MetaCells (Persad et al., 2023), and quantify the gene expression mean and covariance for each. Here too we apply BW-FM conditioned on cell-state, which encompass the heterogeneity of immune profiles appearing as response to COVID infection. Despite the plurality of labels, BW-FM can synthesize correct examples for each condition; see Figure S3.

4.2. Families of general distributions

Recall that for general distributions, we use entropic optimal transport to estimate McCann interpolations and velocities (Cuturi, 2013). With high-resolution CAD datasets, we samplr 2048-point realizations from each shape to approximate OT maps during training. For spatial transcriptomics, MNIST and Letters datasets, we work directly with their naturally varying sample sizes, leveraging entropic OT's ability to handle unequal point-cloud sizes while maintaining accurate transport estimation. These align with our assumption that the underlying distributions are continuous, with point-clouds serving as discrete realizations.

	Airplane		Chair		Car	
	CD ↓	EMD ↓	CD ↓	EMD ↓	CD ↓	EMD ↓
PointFlow	75.68	70.74	62.84	60.57	58.10	56.25
SoftFlow	76.05	65.80	59.21	60.05	64.77	60.09
DPF-Net	75.18	65.55	62.00	58.53	62.35	54.48
Shape-GF	80.00	76.17	68.96	65.48	63.20	56.53
PVD	73.82	64.81	56.26	53.32	54.55	53.83
PSF	71.11	61.09	58.92	54.45	57.19	56.07
WFM (ours)	69.88	64.44	57.62	57.93	53.41	58.10

Table 3. Performance evaluation using 1-NN Accuracy (Earth Mover's Distance and Chamfer Distance) on distributions from ShapeNet, realized as uniform 3D point-clouds. Top three results for each metric are highlighted (1st, 2nd, 3rd). WFM achieves state-of-the-art CD performance on 2 datasets while maintaining competitive results across other metrics (data from Wu et al. (2023a)), while producing diverse samples, see Figure 3.

We compare WFM to many other generative models. Following in their footsteps, we measure generation quality based on the 1-Nearest-Neighbour accuracy metric between generated and test-set shapes. On uniform, 3D datasets, WFM is competitive with current approaches, but exemplifies itself with its unique ability to generate samples with varying sizes and in high-dimensions; see Tables 3 & 4.

4.2.1. 2D & 3D Shapes

Derived from 3D CAD designs, ShapeNet & ModelNet (Wu et al., 2015; Chang et al., 2015) are touchstone shape datasets in computational geometry. Trained individually on samples from the *chair*, *car* and *plane* classes of ShapeNet, WFM synthesized high quality shapes with diverse profiles and matches the performance of previous 3D generation algorithms; see Figure 3 and Table 3. Our framework's versatility allows for seamless integration of label information during training, enabling the synthesis of shapes conditioned

on specific classes. On the full 40-class ModelNet dataset, WFM learned condition-dependent flows between distributions, enabling generation of diverse shapes from the same initial distribution based on the desired label; see Figure 3. We stress that WFM is not restricted to only noisy source measures but can generate transformations between any two collections of distributions. To this end, we demonstrate that WFM can interpolate between two arbitrary elements in the dataset (e.g., between a lamp and a handbag) and complete shapes based on partial profiles (e.g., generate the remaining parts of a plane); see Figure S5.

	Current methods		WFM (ours)	
	CD ↓	EMD ↓	CD ↓	EMD ↓
Variable-sized realizations				
MNIST (4)	NA	NA	63.34	59.97
Letters (A)	NA	NA	62.12	58.68
High-dimensional distributions				
MERFISH	NA	NA	52.80	54.52
XENIUM	NA	NA	60.69	64.20

Table 4. 1-Nearest-Neighbour Accuracy demonstrating WFM's unique capabilities. Using transformers and entropic transport maps, WFM handles both variable-sized realizations and high-dimensional distributions, unlike previous approaches limited to fixed-size, low-dimensional samples.

Another novel facet of WFM is its ability to perform generative modeling from inhomogeneous datasets, where the number of points per point-cloud realization varies between independent samples. This happens in the MNIST or Letters datasets, where data is generated by thresholding low-resolution grayscale images. WFM sets itself apart from other methods, which are restricted to uniform datasets, by leveraging the entropic OT map's ability to compute feasible transformations between empirical samples of different sizes. Our experiments in Figure 4 demonstrate that WFM generates high-quality & diverse samples, despite large variability in the number of particles, itself a novel contribution.

4.2.2. Spatial Transcriptomics

In spatial transcriptomics, a cell's niche is represented by the distribution of gene expression in its neighboring cells, which we observe through its immediate nearest neighbors in high-dimensional space. This approach is complementary to the BW representation of a niche (recall Section 4.1.2), and serves as a more high fidelity view suited for fine-grain interactions. Due to their high dimensionality, these cellular microenvironment distributions have remained beyond the reach of generative models that depend on voxel-based neural networks. Instead, WFM uses transformers, whose permutation equivariance and indifference to dimensionality make them natural architectures for modeling these high-dimensional distributions through their point-cloud realizations (Haviv et al., 2024b).

In the motor cortex, somatostatin (*Sst*) interneurons are organized into phenotypically distinct subtypes, each localized to specific cortical layers (Wu et al., 2023b). Using the 254-gene MERFISH atlas (Zhang et al., 2021), we represent each *Sst* neuron's microenvironment as a distribution in gene expression space, observed through its 16 nearest neighbors. Despite operating in this high-dimensional space, WFM successfully generates niche distributions that capture both global tissue organization and subtype-specific molecular signatures; see Figures 5 & 6.

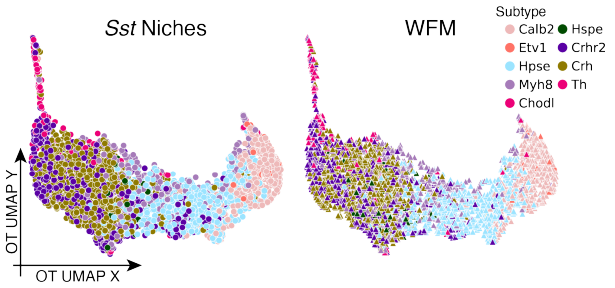

Figure 5. WFM enables high-dimensional generation of cellular microenvironments. (A) Wasserstein 2D embedding of observed *Sst*-neuron niches in motor cortex, colored by subtype. (B) WFM generated niches faithfully reproduce of the tissue landscape

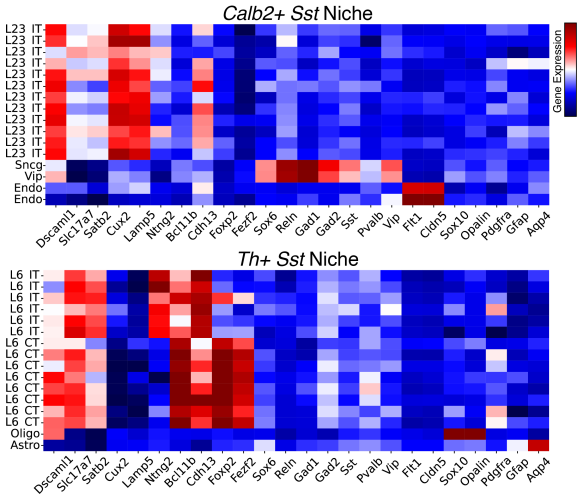

Figure 6. Examples of generated niches for *Calb2*+ and *Th*+ *Sst*-neurons (see Figure 5) display expected gene-expression signatures, with *Calb2*+ niches enriched for upper-layer markers (*Cux2*) and *Th*+ for deep-layer markers (*Ntng2* and *Foxp2*). Each niche is represented as a high-dimensional distribution realized through point-clouds, demonstrating WFM's unique capability to generate distributions in spaces where previous methods, limited by voxelization, cannot operate.

5. Conclusion and outlook

This work shows how to appropriately *lift* the Riemannian flow matching paradigm of Chen & Lipman (2023) to the Wasserstein space, resulting in Wasserstein flow matching. Our motivations stem from modern datasets, where each data sample is a probability distribution, necessitating this

extension for generative modeling purposes. Our contributions are algorithmic in nature, which incorporate various elements, such as estimating optimal transport maps via entropic OT, closed-form expressions over the Bures–Wasserstein space, and attention mechanisms in neural network architectures. Our algorithm is capable of generating realistic data from Gaussian and general distribution realized as variable-size or high-dimensional point-clouds. Both contexts are highly relevant in single-cell transcriptomics for synthesizing of microenvironments and cellular states.

Impact Statement

Wasserstein Flow Matching is designed for generative modeling of distributions, with applications in computational biology. As with any generative model, care should be taken when using WFM to draw biological conclusions.

Acknowledgements

AAP acknowledges support from NSF grant DMS-1922658. DP is an Investigator at the Howard Hughes Medical Institute and is supported by National Cancer Institute grants P30 CA08748 and U54 CA209975.

Meta was involved only in an advisory role. All experimentation and data processing was conducted at MSKCC.

References

Achlioptas, P., Diamanti, O., Mitliagkas, I., and Guibas, L. Learning representations and generative models for 3d point clouds–supplementary material–. arXiv preprint arXiv:1801.07829, 2018. Cited on page 3.

Albergo, M. S. and Vanden-Eijnden, E. Building normalizing flows with stochastic interpolants. arXiv preprint arXiv:2209.15571, 2022. Cited on pages 1 and 3.

Altschuler, J., Weed, J., and Rigollet, P. Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration. In Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA, 2017. Cited on page 13.

Altschuler, J., Chewi, S., Gerber, P. R., and Stromme, A. Averaging on the Bures–Wasserstein manifold: Dimension-free convergence of gradient descent. Advances in Neural Information Processing Systems, 34:22132–22145, 2021. Cited on page 16.

Amari, S.-i. Information geometry and its applications, volume 194. Springer, 2016. Cited on page 3.

Ambrosio, L., Gigli, N., and Savaré, G. Gradient flows in metric spaces and in the space of probability measures. Lectures in Mathematics ETH Zürich. Birkhäuser Verlag, Basel, second edition, 2008. ISBN 978-3-7643-8721-1. Cited on page 3.

Amos, B., Cohen, S., Luise, G., and Redko, I. Meta optimal transport. arXiv preprint arXiv:2206.05262, 2022. Cited on page 18.

Atanackovic, L., Zhang, X., Amos, B., Blanchette, M., Lee, L. J., Bengio, Y., Tong, A., and Neklyudov, K. Meta flow matching: Integrating vector fields on the wasserstein manifold. arXiv preprint arXiv:2408.14608, 2024. Cited on page 3.

Ba, J. L. Layer normalization. arXiv preprint arXiv:1607.06450, 2016. Cited on page 18.

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. Cited on page 4.

Bennett, J. OpenStreetMap. Packt Publishing Ltd, 2010. Cited on page 20.

Binnewies, M., Roberts, E. W., Kersten, K., Chan, V., Fearon, D. F., Merad, M., Coussens, L. M., Gabrilovich, D. I., Ostrand-Rosenberg, S., Hedrick, C. C., et al. Understanding the tumor immune microenvironment (time) for effective therapy. Nature medicine, 24(5):541–550, 2018. Cited on page 2.

Boeing, G. Osmnx: New methods for acquiring, constructing, analyzing, and visualizing complex street networks. Computers, environment and urban systems, 65: 126–139, 2017. Cited on page 20.

Bradbury, J., Frostig, R., Hawkins, P., Johnson, M. J., Leary, C., Maclaurin, D., Necula, G., Paszke, A., VanderPlas, J., Wanderman-Milne, S., et al. Jax: Autograd and xla. Astrophysics Source Code Library, pp. ascl–2111, 2021. Cited on page 18.

Brenier, Y. Polar factorization and monotone rearrangement of vector-valued functions. Comm. Pure Appl. Math., 44 (4):375–417, 1991. Cited on page 3.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y. T., Li, Y., Lundberg, S., et al. Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712, 2023. Cited on page 1.

Cai, R., Yang, G., Averbuch-Elor, H., Hao, Z., Belongie, S., Snavely, N., and Hariharan, B. Learning gradient fields for shape generation. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part III 16, pp. 364–381. Springer, 2020. Cited on page 3.

Carlen, E. A. and Gangbo, W. Constrained steepest descent in the 2-Wasserstein metric. Annals of mathematics, pp. 807–846, 2003. Cited on page 17.

Chang, A. X., Funkhouser, T., Guibas, L., Hanrahan, P., Huang, Q., Li, Z., Savarese, S., Savva, M., Song, S., Su, H., et al. Shapenet: An information-rich 3d model repository. arXiv preprint arXiv:1512.03012, 2015. Cited on page 7.

Chen, R. T. and Lipman, Y. Riemannian flow matching on general geometries. arXiv preprint arXiv:2302.03660, 2023. Cited on pages 1, 2, 3, 5, 8, 15, 16, and 17.

Chen, Y., Goldstein, M., Hua, M., Albergo, M. S., Boffi, N. M., and Vanden-Eijnden, E. Probabilistic forecasting with stochastic interpolants and Föllmer processes. arXiv preprint arXiv:2403.13724, 2024. Cited on page 5.

Cheng, C., Li, J., Peng, J., and Liu, G. Categorical flow matching on statistical manifolds. arXiv preprint arXiv:2405.16441, 2024. Cited on page 3.

Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., and Wang, B. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. Nature Methods, pp. 1–11, 2024. Cited on page 3.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. Advances in Neural Information Processing Systems, 26, 2013. Cited on pages 6 and 7.

Cuturi, M., Meng-Papaxanthos, L., Tian, Y., Bunne, C., Davis, G., and Teboul, O. Optimal transport tools (OTT): A JAX toolbox for all things Wasserstein. arXiv preprint arXiv:2201.12324, 2022. Cited on pages 5, 13, and 18.

Davis, O., Kessler, S., Petrache, M., Bose, A. J., et al. Fisher flow matching for generative modeling over discrete data. arXiv preprint arXiv:2405.14664, 2024. Cited on page 3.

Deb, N., Ghosal, P., and Sen, B. Rates of estimation of optimal transport maps using plug-in estimators via barycentric projections. Advances in Neural Information Processing Systems, 34:29736–29753, 2021. Cited on page 14.

Divol, V., Niles-Weed, J., and Pooladian, A.-A. Optimal transport map estimation in general function spaces. arXiv preprint arXiv:2212.03722, 2022. Cited on pages 5 and 14.

Emami, P. and Pass, B. Optimal transport with optimal transport cost: the Monge–Kantorovich problem on Wasserstein spaces. Calculus of Variations and Partial Differential Equations, 64(2):43, 2025. Cited on page 16.

Eyring, L., Klein, D., Uscidda, T., Palla, G., Kilbertus, N., Akata, Z., and Theis, F. Unbalancedness in neural monge maps improves unpaired domain translation, 2024. URL https://arxiv.org/abs/2311.15100. Cited on page 3.

Flamary, R., Courty, N., Gramfort, A., et al. Pot: Python optimal transport. Journal of Machine Learning Research, 22(78):1–8, 2021. URL http://jmlr.org/papers/v22/20-451.html. Cited on pages 5 and 13.

Gadelha, M., Wang, R., and Maji, S. Multiresolution tree networks for 3d point cloud processing. In Proceedings of the European Conference on Computer Vision (ECCV), pp. 103–118, 2018. Cited on page 3.

Gayoso, A., Lopez, R., Xing, G., Boyeau, P., Valiollah Pour Amiri, V., Hong, J., Wu, K., Jayasuriya, M., Mehlman, E., Langevin, M., et al. A python library for probabilistic analysis of single-cell omics data. Nature biotechnology, 40(2):163–166, 2022. Cited on page 3.

Gelbrich, M. On a formula for the l2 Wasserstein metric between measures on Euclidean and Hilbert spaces. Mathematische Nachrichten, 147(1):185–203, 1990. Cited on page 4.

Haviv, D., Kunes, R. Z., Dougherty, T., Burdziak, C., Nawy, T., Gilbert, A., and Pe'Er, D. Wasserstein wormhole: Scalable optimal transport distance with transformers. ArXiv, 2024a. Cited on page 20.

Haviv, D., Remšík, J., Gatie, M., Snopkowski, C., et al. The covariance environment defines cellular niches for spatial inference. Nature Biotechnology, pp. 1–12, 2024b. Cited on pages 1, 2, 7, 8, 17, and 19.

Hütter, J.-C. and Rigollet, P. Minimax estimation of smooth optimal transport maps. The Annals of Statistics, 49(2): 1166–1194, 2021. Cited on pages 5 and 14.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. Highly accurate protein structure prediction with alphafold. nature, 596(7873):583–589, 2021. Cited on page 1.

Kim, H., Lee, H., Kang, W. H., Lee, J. Y., and Kim, N. S. Softflow: Probabilistic framework for normalizing flow on manifolds. Advances in Neural Information Processing Systems, 33:16388–16397, 2020. Cited on page 3.

Kingma, D. P. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014. Cited on page 18.

Klein, D., Uscidda, T., Theis, F., and Cuturi, M. Genot: Entropic (gromov) wasserstein flow matching with applications to single-cell genomics, 2024. URL https://arxiv.org/abs/2310.09254. Cited on page 3.

Klokov, R., Boyer, E., and Verbeek, J. Discrete point flow networks for efficient point cloud generation. In European Conference on Computer Vision, pp. 694–710. Springer, 2020. Cited on page 3.

Lambert, M., Chewi, S., Bach, F., Bonnabel, S., and Rigollet, P. Variational inference via Wasserstein gradient flows. Advances in Neural Information Processing Systems, 35: 14434–14447, 2022. Cited on page 4.

Lee, J., Lee, Y., Kim, J., Kosiorek, A., Choi, S., and Teh, Y. W. Set transformer: A framework for attention-based permutation-invariant neural networks. In International conference on machine learning, pp. 3744–3753. PMLR, 2019. Cited on page 2.

Lipman, Y., Chen, R. T., Ben-Hamu, H., Nickel, M., and Le, M. Flow matching for generative modeling. arXiv preprint arXiv:2210.02747, 2022. Cited on pages 1, 3, and 15.

Liu, X., Gong, C., and Liu, Q. Flow straight and fast: Learning to generate and transfer data with rectified flow. arXiv preprint arXiv:2209.03003, 2022. Cited on pages 1 and 3.

Lohoff, T., Ghazanfar, S., Missarova, A., Koulena, N., Pierson, N., Griffiths, J. A., Bardot, E. S., Eng, C.-H., Tyser, R. C., Argelaguet, R., et al. Integration of spatial and single-cell transcriptomic data elucidates mouse organogenesis. Nature biotechnology, 40(1):74–85, 2022. Cited on pages 7 and 19.

Lopez, R., Regier, J., Cole, M. B., Jordan, M. I., and Yosef, N. Deep generative modeling for single-cell transcriptomics. Nature methods, 15(12):1053–1058, 2018. Cited on pages 1 and 3.

Manole, T., Balakrishnan, S., Niles-Weed, J., and Wasserman, L. Plugin estimation of smooth optimal transport maps. arXiv preprint arXiv:2107.12364, 2021. Cited on page 5.

Mena, G. and Niles-Weed, J. Statistical bounds for entropic optimal transport: sample complexity and the central limit theorem. Advances in Neural Information Processing Systems, 32, 2019. Cited on page 13.

Muzellec, B., Vacher, A., Bach, F., Vialard, F.-X., and Rudi, A. Near-optimal estimation of smooth transport maps with kernel sums-of-squares. arXiv preprint arXiv:2112.01907, 2021. Cited on page 14.

Nielsen, F. An elementary introduction to information geometry. Entropy, 22(10):1100, 2020. Cited on page 3.

Nowotschin, S., Setty, M., Kuo, Y.-Y., Liu, V., Garg, V., Sharma, R., Simon, C. S., Saiz, N., Gardner, R., Boutet, S. C., et al. The emergent landscape of the mouse gut endoderm at single-cell resolution. Nature, 569(7756): 361–367, 2019. Cited on page 7.

OpenAI. Dall.e-2. https://openai.com/dall-e-2/, 2022. Cited on page 1.

Otto, F. The geometry of dissipative evolution equations: The porous medium equation. 2001. Cited on page 3.

Panaretos, V. M. and Zemel, Y. The Wasserstein Space, pp. 37–57. Springer International Publishing, Cham, 2020. ISBN 978-3-030-38438-8. Cited on page 14.

Persad, S., Choo, Z.-N., Dien, C., Sohail, N., Masilionis, I., et al. Seacells infers transcriptional and epigenomic cellular states from single-cell genomics data. Nature Biotechnology, 41(12):1746–1757, 2023. Cited on pages 1, 2, 7, and 19.

Peyré, G. and Cuturi, M. Computational optimal transport. Foundations and Trends® in Machine Learning, 11(5-6): 355–607, 2019. Cited on page 13.

Pooladian, A.-A. and Niles-Weed, J. Entropic estimation of optimal transport maps. arXiv preprint arXiv:2109.12004, 2021. Cited on pages 2, 5, 13, and 14.

Pooladian, A.-A., Cuturi, M., and Niles-Weed, J. Debiaser beware: Pitfalls of centering regularized transport maps. arXiv preprint arXiv:2202.08919, 2022. Cited on page 14.

Pooladian, A.-A., Ben-Hamu, H., Domingo-Enrich, C., Amos, B., Lipman, Y., and Chen, R. T. Multisample flow matching: Straightening flows with minibatch couplings. arXiv preprint arXiv:2304.14772, 2023a. Cited on page 17.

Pooladian, A.-A., Divol, V., and Niles-Weed, J. Minimax estimation of discontinuous optimal transport maps: The semi-discrete case. arXiv preprint arXiv:2301.11302, 2023b. Cited on page 14.

Rigollet, P. and Stromme, A. J. On the sample complexity of entropic optimal transport. arXiv preprint arXiv:2206.13472, 2022. Cited on page 14.

Santambrogio, F. Optimal transport for applied mathematicians. Birkäuser, NY, 55(58-63):94, 2015. Cited on page 16.

Scetbon, M., Cuturi, M., and Peyré, G. Low-rank sinkhorn factorization. In International Conference on Machine Learning, pp. 9344–9354. PMLR, 2021. Cited on page 18.

Sinkhorn, R. A relationship between arbitrary positive matrices and doubly stochastic matrices. The annals of mathematical statistics, 35(2):876–879, 1964. Cited on pages 6 and 13.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations. arXiv preprint arXiv:2011.13456, 2020. Cited on page 1.

Stark, H., Jing, B., Wang, C., Corso, G., Berger, B., Barzilay, R., and Jaakkola, T. Dirichlet flow matching with applications to DNA sequence design. arXiv preprint arXiv:2402.05841, 2024. Cited on page 3.

Stephenson, E., Reynolds, G., Botting, R. A., Calero-Nieto, F. J., Morgan, M. D., Tuong, Z. K., Bach, K., Sungnak, W., Worlock, K. B., Yoshida, M., et al. Single-cell multi-omics analysis of the immune response in covid-19. Nature medicine, 27(5):904–916, 2021. Cited on page 7.

Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., Mantineo, H., Brydon, E. M., Zeng, Z., Liu, X. S., et al. Transfer learning enables predictions in network biology. Nature, 618(7965):616–624, 2023. Cited on page 3.

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. Cited on page 17.

Vaswani, A. Attention is all you need. Advances in Neural Information Processing Systems, 2017. Cited on pages 2 and 18.

Villani, C. Optimal transport: old and new, volume 338. Springer, 2009. Cited on pages 3 and 16.

Vyas, A., Shi, B., Le, M., Tjandra, A., Wu, Y.-C., Guo, B., Zhang, J., Zhang, X., Adkins, R., Ngan, W., et al. Audiobox: Unified audio generation with natural language prompts. arXiv preprint arXiv:2312.15821, 2023. Cited on page 1.

Wang, W., Ozolek, J. A., Slepčev, D., Lee, A. B., Chen, C., and Rohde, G. K. An optimal transportation approach for nuclear structure-based pathology. IEEE transactions on medical imaging, 30(3):621–631, 2010. Cited on page 14.

Wu, L., Wang, D., Gong, C., Liu, X., Xiong, Y., Ranjan, R., Krishnamoorthi, R., Chandra, V., and Liu, Q. Fast point cloud generation with straight flows. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 9445–9454, 2023a. Cited on pages 3, 7, and 20.

Wu, S. J., Sevier, E., Dwivedi, D., Saldi, G.-A., Hairston, A., Yu, S., Abbott, L., Choi, D. H., Sherer, M., Qiu, Y., et al. Cortical somatostatin interneuron subtypes form cell-type-specific circuits. Neuron, 111(17):2675–2692, 2023b. Cited on page 8.

Wu, Z., Song, S., Khosla, A., Yu, F., Zhang, L., Tang, X., and Xiao, J. 3d shapenets: A deep representation for volumetric shapes. In Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1912–1920, 2015. Cited on page 7.

Xing, Z., Feng, Q., Chen, H., Dai, Q., Hu, H., Xu, H., Wu, Z., and Jiang, Y.-G. A survey on video diffusion models. ACM Computing Surveys, 2023. Cited on page 1.

Yang, G., Huang, X., Hao, Z., Liu, M.-Y., Belongie, S., and Hariharan, B. Pointflow: 3d point cloud generation with continuous normalizing flows. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 4541–4550, 2019. Cited on page 3.

Zemel, Y. and Panaretos, V. M. Fréchet means and Procrustes analysis in Wasserstein space. 2019. Cited on page 16.

Zhang, L., Tozzo, V., Higgins, J., and Ranganath, R. Set norm and equivariant skip connections: Putting the deep in deep sets. In International Conference on Machine Learning, pp. 26559–26574. PMLR, 2022. Cited on page 18.

Zhang, M., Eichhorn, S. W., Zingg, B., Yao, Z., Cotter, K., Zeng, H., Dong, H., and Zhuang, X. Spatially resolved cell atlas of the mouse primary motor cortex by merfish. Nature, 598(7879):137–143, 2021. Cited on page 8.

Zhou, L., Du, Y., and Wu, J. 3d shape generation and completion through point-voxel diffusion. In Proceedings of the IEEE/CVF international conference on computer vision, pp. 5826–5835, 2021. Cited on page 3.

A. Entropic estimation of OT maps

We briefly discuss how to estimate optimal transport maps between distributions using entropic OT. We refer the interested reader to Pooladian & Niles-Weed (2021) for more information on this approach.

We first outline the numerical aspects of the approach; we follow Peyré & Cuturi (2019). Let $\mu = \sum_i m^{-1}\delta_{x_i}$ and $\nu = \sum_j n^{-1}\delta_{y_j}$ be two empirical distributions where $\boldsymbol{X} = \{x_1, \dots, x_m\}$, $\boldsymbol{Y} = \{y_1, \dots, y_n\}$. We first define the following polyhedral constraint set

$$U_{m,n} \coloneqq \left\{P \in \mathbb{R}_+^{m\times n} \,:\, P\mathbf{1}_m = m^{-1}\mathbf{1}_m\,, P^\top \mathbf{1}_n = n^{-1}\mathbf{1}_n\right\},$$

which represents the possible couplings between the two discrete measures. The entropic optimal transport coupling between the two discrete measures μ and ν is defined as the minimizer to the following strictly convex optimization problem

$$\boldsymbol{P}^\star \coloneqq \operatorname*{argmin}_{P\in U_{m,n}} \langle C, P\rangle + \varepsilon H(P)\,, \tag{12}$$

where $\varepsilon > 0$, $H(P) \coloneqq \sum_{i,j} P_{i,j}(\log(P_{i,j}) - 1)$, and $C_{i,j} \coloneqq \|x_i - y_j\|_2^2$. Sinkhorn's matrix scaling algorithm (Sinkhorn, 1964) makes it possible to solve for $\boldsymbol{P}^\star$ with a runtime of $O(mn/\varepsilon)$ (Altschuler et al., 2017). We briefly stress three points:

1. The coupling $\boldsymbol{P}^\star$ is *not* a permutation matrix. The coupling lies inside the polytope $U_{m,n}$ and not at the vertices, and therefore is not a permutation matrix.
2. When $\varepsilon = 0$, the objective becomes a standard linear program with a runtime of $\tilde{O}(mn(m+n))$ (up to log factors) (Peyré & Cuturi, 2019, Chapter 3). While we include a CPU implementation (Flamary et al., 2021) in the WFM codebase, this approach lacks GPU efficiency and substantially increases training time, making it impractical for most use cases.
3. Instead, the regularization parameter ε serves as a tunable training hyperparameter. Lower ε values better approximate true the optimal transport map but require more Sinkhorn iterations for convergence, creating a direct trade-off between accuracy and computational efficiency.

In all our experiments, we used the open-source package OTT-JAX (See https://ott-jax.readthedocs.io/en/latest/) to compute the entropic coupling and the out-of-sample mapping (Cuturi et al., 2022).

A.1. Rounded matchings

Our first approach holds when $m = n$. In this case, we can greedily *round* the noisy matching matrix $\boldsymbol{P}^\star$ to become a permutation. This is achieved through an iterative process of selecting the maximum value (argmax) and zeroing out corresponding rows and columns. This method repeatedly identifies the largest remaining probability, sets it to 1, and eliminates other entries in its row and column, ultimately resulting in a permutation matrix that preserves the probabilistic assignment implied by the original doubly stochastic matrix. This is merely a GPU-friendly heuristic approximation to the true optimal permutation matrix between the two point-clouds.

A.2. Entropic transport map: An out-of-sample estimator

A primal-dual relationship of the strictly convex program (12) shows that there exist vectors $(\boldsymbol{f}^\star, \boldsymbol{g}^\star) \in \mathbb{R}^m \times \mathbb{R}^n$ such that

$$\boldsymbol{P}^\star_{i,j} = e^{\boldsymbol{f}_i^\star/\varepsilon} e^{-\boldsymbol{C}_{i,j}/\varepsilon} e^{\boldsymbol{g}_j^\star/\varepsilon}$$

These two vectors are called the Kantorovich potentials, which are initially defined on the support of μ and ν, respectively. However, they can be readily extended to all of $\mathbb{R}^d$ (Mena & Niles-Weed, 2019), resulting in two functions

$$\hat{f}(x) = -\varepsilon \log\Big(\sum_{j=1}^n n^{-1} \exp((\boldsymbol{g}_j^\star - \|x - y_j\|^2)/\varepsilon)\Big)\,,$$

$$\hat{g}(y) = -\varepsilon \log\Big(\sum_{i=1}^m m^{-1} \exp((\boldsymbol{f}_i^\star - \|y - x_i\|^2)/\varepsilon)\Big)\,.$$

Following Pooladian & Niles-Weed (2021), we can define the *entropic transport map*, where the last equality is a simple calculation:

$$\hat{T}_\varepsilon(x) \coloneqq x - \nabla \hat{f}(x) = \frac{\sum_{j=1}^n y_j \exp((\boldsymbol{g}_j^\star - \|x - y_j\|^2)/\varepsilon)}{\sum_{j=1}^n \exp((\boldsymbol{g}_j^\star - \|x - y_j\|^2)/\varepsilon)} \,. \tag{13}$$

This estimator was initially to provide statistical approximations to the optimal transport map $T_\star^{\mu\to\nu}$ on the basis of samples; see Pooladian & Niles-Weed (2021); Rigollet & Stromme (2022); Pooladian et al. (2023b; 2022). Note that $\hat{T}_\varepsilon(x)$ can be interpreted as the conditional expectation of the plan $\boldsymbol{P}^\star$ conditioned on out-of-sample inputs $x \in \mathbb{R}^d$, which is well-defined due to the relations above. Finally, we stress that this estimator can be adapted to settings where the point-clouds μ and ν not only have different numbers of points, but also non-uniform weights. As this estimator is also a by-product of Sinkhorn's algorithm, it is also scalable and GPU-friendly.

A.3. On the (statistical) approximations of geodesics

We briefly collect a basic results pertaining to the (statistical) approximation of optimal transport paths. This bound shows that the error grows along the trajectory, but is limited by the overall distance of the maps.

Proposition A.1. *Let μ, ν be two probability measures and suppose μ has a density, and let $T^\star$ be the optimal transport map from μ to ν. Let $\hat{T}$ be an estimator to the optimal transport map, defined with respect to data $X_1, \dots, X_n \sim \mu$ and $Y_1, \dots, Y_n \sim \nu$. Then for $t \in [0,1]$*

$$\mathbb{E}[W_2^2(\rho_t, \hat{\rho}_t)] \le t^2 \mathbb{E}\|\hat{T} - T^\star\|_{L^2(\mu)}^2 \,,$$

where the outer expectation is taken with respect to the data, and we define

$$\rho_t \coloneqq ((1-t)\mathrm{id} + tT^\star)_\sharp \mu \,, \quad \hat{\rho}_t \coloneqq ((1-t)\mathrm{id} + t\hat{T})_\sharp \mu$$

Proof. The result follows immediately from a standard coupling argument to obtain the linearized Wasserstein distance (Wang et al., 2010; Panaretos & Zemel, 2020)

$$W_2^2(\rho_t, \hat{\rho}_t) \le \|((1-t)\mathrm{id} + tT^\star) - ((1-t)\mathrm{id} + t\hat{T})\|_{L^2(\mu)}^2 = t^2 \|\hat{T} - T^\star\|_{L^2(\mu)}^2 \,. \qquad \square$$

The entropic Brenier map is one particular estimator. We note two key properties of this map; see Pooladian & Niles-Weed (2021) for in-depth discussions.

Theorem A.2. *Suppose μ, ν have density bounded above and below, and that the optimal transport map between them, denoted $T^\star$, is such that $(T^\star)^{-1}$ is at least twice differentiable and there exists $\lambda, \Lambda > 0$ such that*

$$\lambda I \preceq DT^\star \preceq \Lambda I \,.$$

Then, when estimated from n samples from μ and n samples from μ, the entropic Brenier map has the following error

$$\mathbb{E}\|\hat{T}_\varepsilon - T^\star\|_{L^2(\mu)}^2 \lesssim n^{-1/2} \log(n) \varepsilon^{-d/2-1} + \varepsilon^2 \,, \tag{14}$$

where we suppress constants that depend on our assumptions. Performing a bias-variance trade-off in the regularization parameter, one obtains $\varepsilon = \varepsilon(n) \asymp n^{-1/(d+4)}$ and the total error becomes

$$\mathbb{E}\|\hat{T}_\varepsilon - T^\star\|_{L^2(\mu)}^2 \lesssim_{\log(n)} n^{-2/(d+4)} \,.$$

We emphasize that the assumptions in Theorem A.2 are standard in the literature (Hütter & Rigollet, 2021; Deb et al., 2021; Muzellec et al., 2021; Divol et al., 2022). While the rate scales exponentially poorly with the dimension, we stress that existing lower bounds of estimation (see Hütter & Rigollet (2021)) also suffer from the curse of dimensionality, which is unavoidable for this task. Combining these two results, we can compare the geodesic given by the OT map, and the one induced by using the entropic map, where we write

$$\hat{\rho}_t^\varepsilon \coloneqq ((1-t)\mathrm{id} + \mathrm{t}\hat{\mathrm{T}}_\varepsilon)_\sharp \mu \,.$$

Corollary A.3. *Consider the same setting as Theorem A.2. Then the geodesic given by the estimated entropic Brenier map, denoted by $\hat{\rho}_t^\varepsilon$, on the basis of n samples and $\varepsilon \asymp n^{-1/(d+4)}$, is close to the true geodesic with the following error*

$$\mathbb{E}[W_2^2(\rho_t, \hat{\rho}_t^\varepsilon)] \lesssim t^2 n^{-2/(d+4)} \,.$$

B. Derivation of the WFM objective

In this section, we give validity to the WFM for optimization purposes, and our choice of curves. For instance, recall the original Flow Matching objective (Lipman et al., 2022)

$$\mathcal{L}_{\mathrm{FM}}(\theta) \coloneqq \int_0^1 \mathbb{E}_{X_t \sim p_t} \|f_\theta(X_t, t) - u_t(X_t)\|^2 \, \mathrm{d}t \,, \tag{15}$$

where $f_\theta : \mathbb{R}^d \times [0,1] \to \mathbb{R}^d$ is a neural network, and $(p_t, u_t)_{t\in[0,1]}$ are a density-vector field pair that satisfy the continuity equation between two distribution μ and ν.

$$\partial p_t + \nabla \cdot (p_t u_t) = 0 \,, \quad p_0 = \mu \,, p_1 = \nu \,.$$

Note that, implicit in $\mathcal{L}_{\mathrm{FM}}(\theta)$ is the endpoint constraints μ and ν. Now, we average over possibly choices of $\mu \sim \mathfrak{p}_0$ and $\nu \sim \mathfrak{p}_1$, resulting in

$$\int_0^1 \iint \mathbb{E}_{X_t \sim p_t} \|f_\theta(X_t, t) - u_t(X_t)\|^2 \, \mathrm{d}\mathfrak{p}_0(\mu) \, \mathrm{d}\mathfrak{p}_1(\nu) \, \mathrm{d}t \,. \tag{16}$$

As a particular case, take $(p_t, u_t) \leftarrow (\mu_t, v_t)$, where the first argument is the McCann interpolation between μ and ν, and v_t is the optimal velocity field, which is a function of the optimal transport map from μ to ν (recall Section 2.3). This yields our final objective (11), which we recall here for convenience

$$\mathcal{L}_{\mathrm{WFM}}(\theta) \coloneqq \int_0^1 \iint \mathbb{E}_{X_t \sim \mu_t} \|f_\theta(X_t, t) - v_t(X_t)\|^2 \, \mathrm{d}\mathfrak{p}_0(\mu) \, \mathrm{d}\mathfrak{p}_1(\nu) \, \mathrm{d}t \,. \tag{17}$$

When $\mathfrak{p}_0, \mathfrak{p}_1$ are distributions over Gaussians, we have closed-form expressions for all objects of interest. When $\mathfrak{p}_0, \mathfrak{p}_1$ are distributions over general distributions, we approximate the geodesics based on point-cloud samples using entropic Brenier maps and their respective interpolations. We emphasize that the rounded-matching we employ (see Appendix A.1) is also a valid curve.

B.1. Validity of Conditional Flows

The key idea behind Flow Matching is that the (tractable) conditional velocity probability paths come together to form correct marginal velocities and paths from source distribution to target. In Riemannian Flow Matching, given a source and target $\mathfrak{p}_1, \mathfrak{p}_0 \in \mathcal{P}(\mathcal{M})$ distributions over *Riemannian* manifold $\mathcal{M}$, the conditional velocity field from $\mathfrak{p}_0$ to a sample $\nu \sim \mathfrak{p}_1$ is generated by flowing from each point $\mu \sim \mathfrak{p}_0$ towards ν using the geodesic path between them. This produces the conditional probability path $\mathfrak{p}_t(\cdot|\nu)$ generated by the vector field $\mathfrak{u}_t(\cdot|\nu)$, which obey the continuity equation for *Riemannian* manifolds:

$$\partial_t \mathfrak{p}_t(\cdot|\nu) + \mathrm{div}_{\mathcal{M}}(\mathfrak{p}_t(\cdot|\nu)\mathfrak{u}_t(\cdot|\nu)) = 0 \tag{18}$$

where $\mathrm{div}_{\mathcal{M}}$ is the Riemannian divergence operator. Following Theorem 1 in Lipman et al. (2022) and equation 7 in Chen & Lipman (2023), the marginal vector field:

$$\mathfrak{u}_t(\mu_t) = \int_{\mathcal{M}} \mathfrak{u}_t(\mu_t|\nu) \frac{\mathfrak{p}_t(\mu_t|\nu)}{\mathfrak{p}_t(\mu_t)} \, \mathrm{d}\mathfrak{p}_1(\nu) \tag{19}$$

generates the marginal probability path:

$$\mathfrak{p}_t(\mu_t) = \int_{\mathcal{M}} \mathfrak{p}_t(\mu_t|\nu) \, \mathrm{d}\mathfrak{p}_1(\nu) \tag{20}$$

as $\mathfrak{p}_t(\mu_t)$ and $\mathfrak{u}_t(\mu_t)$ together obey the continuity equation. The proof requires regularity constraints on $\mathfrak{p}_t(\mu_t|\nu)$ and $\mathfrak{u}_t(\mu_t|\nu)$, which we assume as in Chen & Lipman (2023). A withstanding requirement that we need is that a unique optimal transport map exists between *any* two measures $(\mu, \nu) \in \mathfrak{p}_0 \times \mathfrak{p}_1$. This can be ensured if, for example, all measures in $\mathfrak{p}_0$

have a density, which is assumed throughout the text. Positivity of the path μ_t can be ensured through a myriad of conditions (e.g., source and target measures have positive density over the same support). In the explicit Gaussian-to-Gaussian case (i.e., the Bures–Wasserstein manifold), this assumption is trivially satisfied as the manifold is a *genuine* finite-dimensional Riemannian manifold and the Riemannian machinery of Chen & Lipman (2023) goes through with no modifications. We carry out the rest of our computations for general distributions over the Wasserstein space, and hence the computations are non-rigorous. For a fully rigorous treatment, see Emami & Pass (2025).

Conditional flows and paths for WFM

To continue, we require the following proposition. Recall that a pre-metric is a function $\mathrm{d} : \mathcal{M} \times \mathcal{M} \to \mathbb{R}$ that satisfies the following three requirements

1. $\mathrm{d}(x, y) \geq 0$ for all $x, y \in \mathcal{M}$,
2. $\mathrm{d}(x, y) = 0$ if and only if $x = y$,
3. $\nabla_x \mathrm{d}(x, y) \neq 0$ if and only if $x \neq y$.

Lemma B.1. *The 2-Wasserstein distance defines a valid pre-metric over $\mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$.*

Proof. As the 2-Wasserstein distance is a metric (Villani, 2009; Santambrogio, 2015), the first and second conditions of a pre-metric are trivially satisfied. For the third condition, we note that for $\mu, \nu \in \mathcal{P}_{2,\mathrm{ac}}(\mathbb{R}^d)$

$$\nabla \tfrac{1}{2} W_2^2(\cdot, \nu)(\mu) = -\log_\mu(\nu) = -2(T^{\mu\to\nu} - \mathrm{id}) \tag{21}$$

For a proof of this computation, see Altschuler et al. (2021) or Zemel & Panaretos (2019). By chain-rule, we have that

$$\nabla \tfrac{1}{2} W_2^2(\cdot, \nu)(\mu) = W_2(\mu, \nu) \nabla W_2(\cdot, \nu)(\mu) = -\log_\mu(\nu)$$

This latter quantity is only zero if and only if $\mu = \nu$, and thus the proof is complete. □

We now recall how to construct conditional vector fields as a function of pre-metrics d, as described by Eq. (13) by (Chen & Lipman, 2023). The construction is the following

$$u_t(x_t|y) = (\tfrac{\mathrm{d}}{\mathrm{d}t} \ln(1-t)) \mathrm{d}(x_t, y) \nabla_x \mathrm{d}(x_t, y) / \|\nabla_x \mathrm{d}(x_t, y)\|^2_{g(x_t)}\,, \tag{22}$$

where g is the metric on the tangent space at x_t, and x_t is a geodesic in the form

$$x_t = \exp_{x_0}((1-t)\log_{x_0}(x_1))\,.$$

Note that our geodesics are precisely of the form

$$\mu_t = \exp_\mu((1-t)\log_\mu(\nu))\,,$$

where we recall $v \mapsto \exp_\mu(v) = (\mathrm{id} + v)_\sharp \mu$ and $\nu \mapsto \log_\mu(\nu) = T^{\mu\to\nu} - \mathrm{id}$. We now specialize these calculations to the case where the pre-metric is the 2-Wasserstein distance.

Lemma B.2. *Fix $\nu \sim \mathfrak{p}_1$. Let $\mu_t \mapsto \mathfrak{u}_t(\mu_t|\nu)$ denote the conditional vector field defined using Equation (22) with d replaced by the 2-Wasserstein distance, evaluated at a geodesic μ_t between any $\mu \sim \mathfrak{p}_0$ and the fixed ν. Then it holds that $\mathfrak{u}_t(\mu_t|\nu) = v_t$, where v_t is given by Equation (7) between μ and ν.*

Proof. We compute, using Equation (21),

$$\begin{aligned}
\mathfrak{u}_t(\mu_t|\nu) &= (\tfrac{\mathrm{d}}{\mathrm{d}t} \ln(1-t)) W_2(\mu_t, \nu) \nabla W_2(\mu_t, \nu) / \|\nabla W_2(\mu_t, \nu)\|^2_{L^2(\mu_t)} \\
&= \frac{1}{1-t}(T^{\mu_t\to\nu} - \mathrm{id}) \\
&= \frac{-1}{1-t}(T^{\mu\to\nu} \circ (T_t)^{-1} - \mathrm{id}) \\
&= \frac{1}{1-t}(T^{\mu\to\nu} - T_t) \circ (T_t)^{-1}\,,
\end{aligned}$$

where we make use of the fact that $T^{\mu_t \to \nu} = T^{\mu \to \nu} \circ (T^{\mu_t \to \nu})^{-1}$ (see e.g., Section 2 of Carlen & Gangbo (2003)), and that $T_t = (1-t)\mathrm{id} + tT^{\mu \to \nu}$. Simplifying, we obtain

$$\mathfrak{u}_t(\mu_t|\nu) = \frac{1}{1-t}(T^{\mu \to \nu} - (1-t)\mathrm{id} - tT^{\mu \to \nu}) \circ (T_t)^{-1} = (T^{\mu \to \nu} - \mathrm{id}) \circ (T_t)^{-1}\,, \tag{23}$$

which is precisely the optimal transport vector field v_t stated in (7). □

Together, we arrive at the following corollary, which validates the probability paths generated by our approach.

Corollary B.3. *The conditional flow $\mathfrak{p}_t(\cdot|\nu)$ generated by the conditional vector fields defined by (23) satisfy $\mathfrak{p}_1(\cdot|\nu) = \nu$.*

Proof. This follows immediately from Theorem 3.1 of Chen & Lipman (2023) by passing through the pre-metric construction above. □

C. Multisample Wasserstein Flow Matching

Since optimal transport can be applied on the Wasserstein manifold itself, both WFM and BW-FM can be seamlessly integrated with the multisample FM (MS-FM) framework (Pooladian et al., 2023a; Tong et al., 2023). The core technique behind MS-FM is to use OT to match minibatches from source and target measures during training, rather than relying on random pairings. This has shown to improve learned flows while requiring fewer function evaluations to synthesize new samples. Applying MS-FM requires computing the pairwise distance matrix between source and target batch samples, denoted from $i \in \{1, \dots, \texttt{Bsz}\}$. In the BW-FM setting, given two sets of Gaussians $\{(a_i, A_i)\}_{i=1}^{\texttt{Bsz}}$ and $\{(b_i, B_i)\}_{i=1}^{\texttt{Bsz}}$, their Frechét ($W_2^2$) distance matrix is:

$$C_{i,j} = \|a_i - b_j\|_2^2 + \mathrm{Tr}(A_i + B_j - 2(A_i^{1/2} B_j A_i^{1/2})^{1/2}) \tag{24}$$

We then use entropic OT to approximately solve the assignment problem on C and compute a transport matrix. This is the converted into a one-to-one assignment matrix via rounded matching (see Appendix A.1), ensuring the entire batch is used in training.

For general WFM, applying MS requires computing pairwise OT distance between all source and target samples within a minibatch. For large point-cloud samples, this is exorbitantly expensive, even with Sinkhorn iterations. For an efficient approximate, here too we rely on the Frechét distance, computed between empirical means and covariances of each point-cloud. Computation of the Frechét distance is markedly less resource-intensive than entropic OT, yet is notably correlated with EMD values (see Table 1 in Haviv et al. (2024b)).

D. Choice of source measure

In both WFM and BW-FM, learning flows requires a source measure which is straightforward to sample from. For a source distribution on the space of $\{(m, \Sigma) : m \in \mathbb{R}^d, \Sigma \in \mathbb{S}_{++}^d\}$, we simply sample means and covariance matrices using independent Gaussian and Wishart distributions, respectively. By default, the parameters for the Gaussian component of the source matches the average and standard deviation of the means in the target, while the scale parameter in the Wishart is the barycenter of the data covariance.

To achieve high-quality generation of general distributions, it is essential that the initial (source) measure be diverse, rather than collapsed and degenerate. Indeed, while it is alluring to produce noise point-clouds by sampling particles from a single base distribution, i.e. $X = \{x_i\}_{i=1}^n$, $x_i \sim \mathcal{N}(0, I_d)$, as n grows, the Wasserstein distance between empirical distributions goes to 0. To alleviate this, we draw point-clouds from multivariate Gaussians with a stochastic covariance:

$$\begin{aligned} L &\sim \mathcal{N}(\mu_L, \sigma_L \cdot I) \\ X &= \{x_i\}_{i=1}^n, x_i \sim \mathcal{N}(0, LL^T) \end{aligned}$$

where μ_L & σ_L are the average and standard deviation of the Cholesky factors from the empirical covariances of the target measure point-clouds. This ensures a wider source measure, producing a diverse range of noise point-clouds.

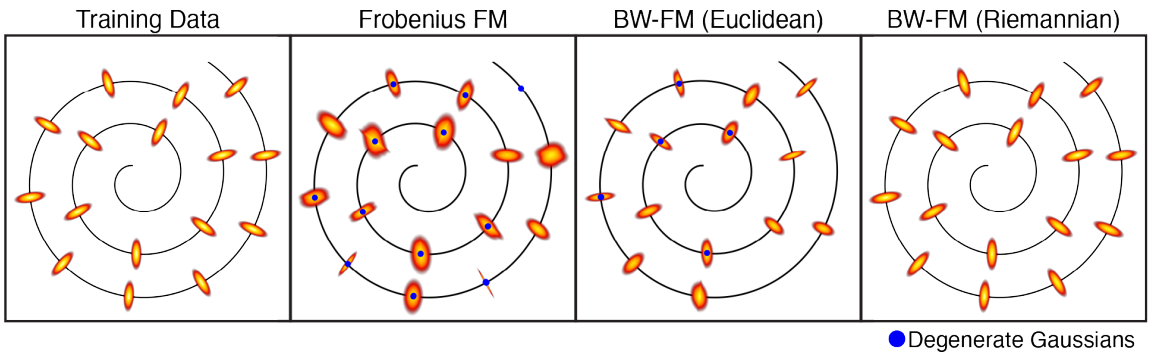

Figure S1. When the number of training examples is too few (in comparison to Figure 2), all methods collapse on the training data, though only the Riemannian instantiation of BW-FM captures the covariances perfectly.

E. Neural architecture & training

E.1. BW-FM on Gaussians

The goal of BW-FM is to train a neural network to match the (Riemannian) time-derivative along the BW geodesics between Gaussians. The model employs a standard, fully connected neural network which takes as input concatenated values of (m_t, Σ_t, t) based on the McCann interpolation formula from Section 2.3.1. Since the covariance matrix is symmetric, only its lower-diagonal values are used, flattened into a vector of length $d(d+1)/2$. Time values are converted to Fourier features, an approach inspired by positional encodings in transformer literature (Vaswani, 2017). To streamline training, two separate networks are employed: one to match the time derivative of the mean $\dot{m}_t$ and another for the time derivative of the covariance matrix $\dot{\Sigma}_t^{\mathrm{BW}}$. The BW tangent norm is used as the loss function for training these networks.

By default, all models use a 6-layer neural network using *relu* non-linearity, with 1024 neurons per layer, applying skip connections and layer-norm (Ba, 2016). Training is performed for $100,000$ gradient descent steps using the Adam optimizer (Kingma, 2014) with an exponential learning rate decay of 0.97 every 1000 steps and batch size of 128.

E.2. WFM on general distributions

WFM estimates the optimal transport (OT) map for a given pair of interpolate point-cloud and time $(\boldsymbol{X}_t, t)$. Here too the time component t is first converted into Fourier features. The model's architecture begins with an embedding layer, followed by a series of alternating multi-head attention and fully-connected layers. Skip connections and layer-norm are applied after each operation. The final layer projects the embeddings back to X's original space using a dense layer with zero initialization. The model is trained by minimizing the squared distance between the predicted and true OT maps.

By default, the entropic OT map is constructed with regularization weight of $\varepsilon = 0.002$ and 200 Sinkhorn iterations, which we found to be sufficient for convergence. Whenever the dataset consists of uniformly sized point-cloud realizations, we use rounded matching (see Appendix A.1), otherwise we apply the out-of-sample estimator (see Appendix A.2) which can calculate maps between point-clouds with different sizes. The transformer network is composed to 6 multi-head attention blocks, with an embedding dimension of 512 and 4 heads. Our model is optimizer with Adam (Kingma, 2014) using an exponential learning rate decay and batch size of 64.

WFM relies on JAX and OTT-JAX (Bradbury et al., 2021; Cuturi et al., 2022) and enjoys seamless optimization via end-to-end just-in-time compilation. For the ShapeNet experiments (see Table 3), the model is trained for 500,000 training steps, totaling to about 3 days of trainings of a single A100 GPU. All other experiments (see Table 4) trained for 100,000 steps, requiring around 3-4 hours of GPU use. We note the Transformer's forward and backwards pass was the most significant source of computational overhead, as opposed to the Sinkhorn based approximation of OT maps. The computational complexity of both components is quadratic with point-cloud size, which limits the scalability of WFM. Making Transformers simpler to optimize and accelerating OT computations are both active areas within ML research (Amos et al., 2022; Scetbon et al., 2021; Zhang et al., 2022), and future iterations of WFM can incorporate solution from those spaces.

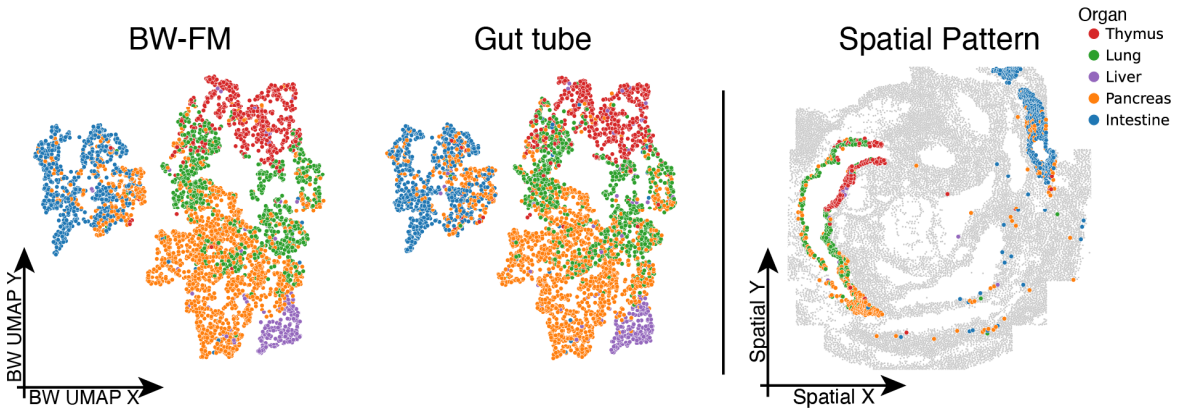

Figure S2. Comparing real and generated cellular microenvironments in embryonic gut tube development. Left: UMAP embedding of BW-FM generated (left panel) and observed (middle panel) microenvironments based on Bures-Wasserstein distance, colored by developing organ type. Right: Spatial arrangement of cells in the seqFISH embryo showing physical organization of organ development.

F. Experiment Details

F.1. Spatial Transcriptomics

In our manuscript, we applied WFM and BW-FM on several spatial transcriptomics datasets, encompassing a variety of technologies and tissue contexts. From the 351-gene seqFISH embryogensis dataset, (Lohoff et al., 2022), we focus on niches of gut-tube cells. We compress gene-expression profiles down to their 16 principal components (PC) and aggregate all the cells around each gut cell within an 80 micron radius, yielding on average 29 cells per niche. We then calculated the gene-expression PC mean and covariance within each environment to produce Gaussians for BW-FM. Generated Gaussians align with real data ad delineate budding organs into their appropriate regions (see Figure S2).

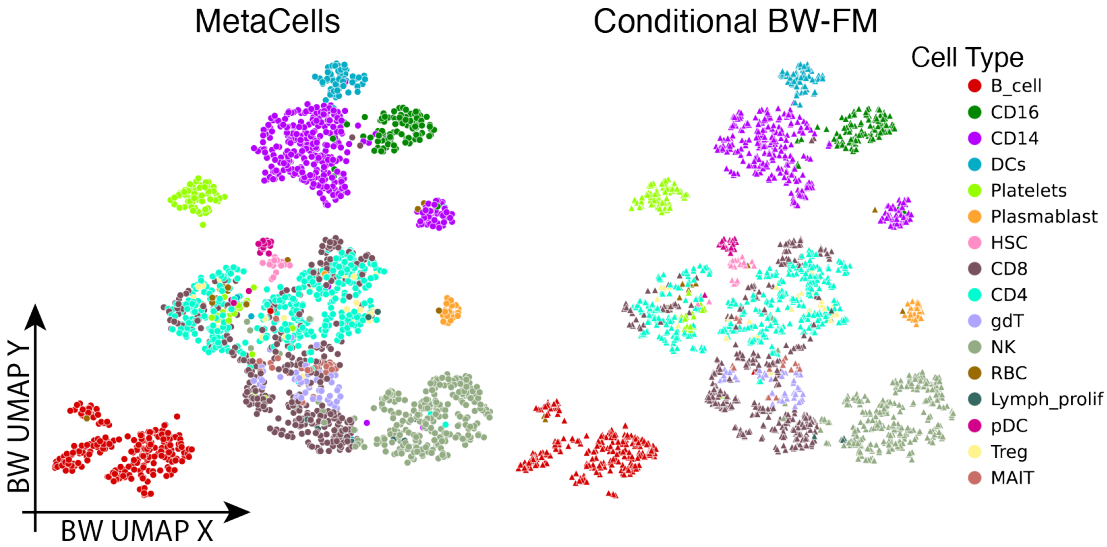

Figure S3. Conditional BW-FM applied to single-cell RNA sequencing data of immune response to COVID-19. Large scale single-cell atlases are commonly grouped into highly dedicated clusters called MetaCells (Persad et al., 2023). In this application, BW-FM is conditioned on cell state and trained to generate means and covariances of gene expression, focusing on the top 32 principal components, derived from aggregated cells. The model achieves high-quality sample generation, as evidenced by a label accuracy of 93.13%.

In a complementary approach, WFM is applied directly on gene-expression based point-clouds of niches, representing a realization of the continuous distribution of their environments. Uniquely suited for high-dimensional data, we apply WFM on a MERFISH atlas of the motor cortex XENIUM assay of melanoma metastasis to the brain (Haviv et al., 2024b). In both dataset, we select the $k = 16$ physical nearest neighbours of every cell, and aggregate expression profiles based on their first

16 PCs.

We concentrated on *Sst* interneurons in the motor cortex datas, which are divided into spatially segregated, phenotypically distinct subtypes. Applied unconditionally, WFM generated niches match the distribution of the real data based on EMD and CD 1-nearest-neighbour accuracy (see Table 4). We then assessed WFM's capability to comprehend the relationship between cell state and environment and tasked it with conditional generation based on subtype label. Based on OT distances estimated via *Wormhole* embeddings (Haviv et al., 2024a), distributions generated by WFM recapitulated true organ environment. The label accuracy for WFM-generated data was 63.86%, which was nearly identical to the test-set real data accuracy of 62.19%.

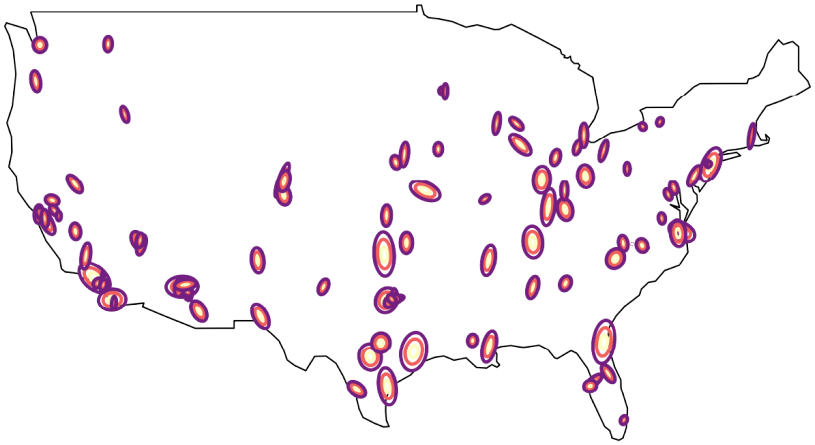

Figure S4. Cities Dataset. Gaussians representing the 100 most populous cities in the continental US. The data was obtained from (Bennett, 2010) via OSMnx (Boeing, 2017). The mean parameter is the longitude and latitude coordinate of each city and the covariance is the 2nd moment approximation of their metro area.

G. 2D & 3D shapes

The ShapeNet dataset consists of 3D shapes of 55 different classes, Emulating the benchmarking effort in (Wu et al., 2023a), we apply WFM to generate $n = 2048$ sized examples from the *plane*, *car* and *chair* classes. At each gradient descent step, we sample 64 shapes from the training set for each class, and randomly select $n = 2048$ points from each. To evaluate generation quality, we synthesize point-clouds to much the size of the test set, and calculate the real or generated $1 - NN$ accuracy based on EMD and CD metrics.

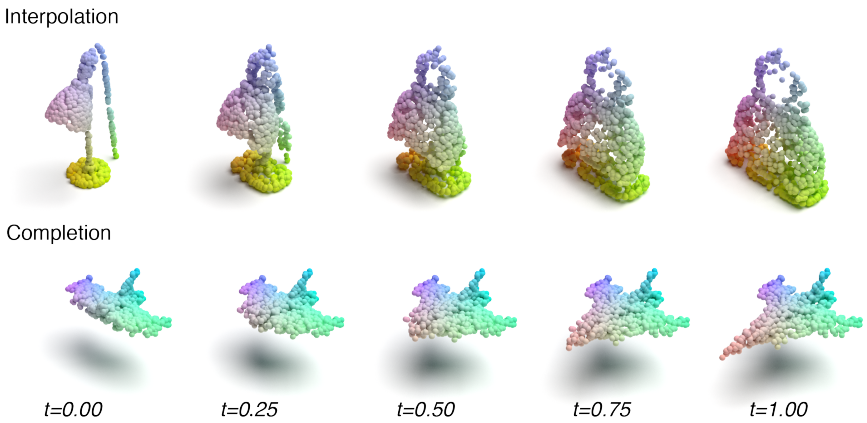

Figure S5. Interpolation and shape completion with WFM. *Top*. Using the *lamps* and *handbags* as the source and target measures, WFM learns to transform a given (unseen test-set) lamp point-cloud into a valid handbag. *Bottom*. Trained to generate full planes, WFM can reconstruct complete point-clouds from partial views of test-set samples.

ModelNet has 40 classes of shapes, each consisting of $n = 2048$ particles. Conditioned on class label, WFM is trained to

generate $n = 1000$ sized point-clouds here too. In this setting, the noise measure is the standard normal and we did not use multi-sample matching. According to nearest-neighbour classification from OT preserving *Wormhole* embeddings, generated samples match their class with an accuracy of 77.66%, approaching the 79.98% purity of test set samples from real-data.

The MNIST dataset is a widely used collection of handwritten digits, consisting of 28x28 pixel grayscale images of the numbers 0 through 9. EMNIST (Extended MNIST) is an expansion of MNIST that includes handwritten letters as well as digits. To convert samples from these datasets into distributions, we threshold each image and extract the coordinates of the above-threshold pixels. This produces a cohort of point-clouds of variables sizes, as each image contains a different number of relevant pixels. We apply the entropic OT map (see Appendix A) based WFM to synthesize distributions of the digit 4 and letter a. Despite the data heterogeneity, WFM produces realistic examples (see Figure 4), while capturing the data distribution (see Table 4).