# Diffusion Earth Mover's Distance and Distribution Embeddings

Alexander Tong [* 1]   Guillaume Huguet [* 2 3]   Amine Natik [* 2 3]   Kincaid MacDonald [4]   Manik Kuchroo [5 1]
Ronald R. Coifman [4]   Guy Wolf [† 2 3]   Smita Krishnaswamy [† 5 1]

## Abstract

We propose a new fast method of measuring distances between large numbers of related high dimensional datasets called the Diffusion Earth Mover's Distance (EMD). We model the datasets as distributions supported on common data graph that is derived from the affinity matrix computed on the combined data. In such cases where the graph is a discretization of an underlying Riemannian closed manifold, we prove that Diffusion EMD is topologically equivalent to the standard EMD with a geodesic ground distance. Diffusion EMD can be computed in $\tilde{O}(n)$ time and is more accurate than similarly fast algorithms such as tree-based EMDs. We also show Diffusion EMD is fully differentiable, making it amenable to future uses in gradient-descent frameworks such as deep neural networks. Finally, we demonstrate an application of Diffusion EMD to single cell data collected from 210 COVID-19 patient samples at Yale New Haven Hospital. Here, Diffusion EMD can derive distances between patients on the manifold of cells at least two orders of magnitude faster than equally accurate methods. This distance matrix between patients can be embedded into a higher level patient manifold which uncovers structure and heterogeneity in patients. More generally, Diffusion EMD is applicable to all datasets that are massively collected in parallel in many medical and biological systems.

## 1. Introduction

With the profusion of modern high dimensional, high throughput data, the next challenge is the integration and analysis of collections of related datasets. Examples of this are particularly prevalent in single cell measurement modalities where data (such as mass cytometry, or single cell RNA sequencing data) can be collected in a multitude of patients, or in thousands of perturbation conditions (Shifrut et al., 2018). These situations motivate the organization and embedding of datasets, similar to how we now organize data points into low dimensional embeddings, e.g., with PHATE (Moon et al., 2019), tSNE (van der Maaten & Hinton, 2008), or diffusion maps (Coifman & Lafon, 2006)). The advantage of such organization is that we can use the datasets as rich high dimensional features to characterize and group the patients or perturbations themselves. In order to extend embedding techniques to entire datasets, we have to define a distance between datasets, which for our purposes are essentially high dimensional point clouds. For this we propose a new form of Earth Mover's Distance (EMD), which we call *Diffusion EMD*[1], where we model the datasets as distributions supported on a common data affinity graph. We provide two extremely fast methods for computing Diffusion EMD based on an approximate multiscale kernel density estimation on a graph.

Optimal transport is uniquely suited to the formulation of distances between entire datasets (each of which is a collection of data points) as it generalizes the notion of the shortest path between two points to the shortest set of paths between distributions. Recent works have applied optimal transport in the single-cell domain to interpolate lineages (Schiebinger et al., 2019; Yang & Uhler, 2019; Tong et al., 2020), interpolate patient states (Tong & Krishnaswamy, 2020), integrate multiple domains (Demetci et al., 2020), or similar to this work build a manifold of perturbations (Chen et al., 2020). All of these approaches use the standard *primal* formulation of the Wasserstein distance. Using either entropic regularization approximation and the Sinkhorn algorithm (Cuturi, 2013) to solve the discrete distribution case or a neural network based approach in the continuous formulation (Ar-

---
[*]Equal contribution ; [†]Equal senior-author contribution. [1]Dept. of Comp. Sci., Yale University, New Haven, CT, USA [2]Dept. of Math. & Stat., Université de Montréal, Montréal, QC, Canada [3]Mila – Quebec AI Institute, Montréal, QC, Canada [4]Dept. of Math., Yale University, New Haven, CT, USA [5]Department of Genetics, Yale University, New Haven, CT, USA. Correspondence to: Smita Krishnaswamy <smita.krishnaswamy@yale.edu>.

---
[1]Python implementation is available at https://github.com/KrishnaswamyLab/DiffusionEMD.

jovsky et al., 2017). We will instead use the *dual* formulation through the well-known Kantorovich-Rubinstein dual to efficiently compute optimal transport between many distributions lying on a common low-dimensional manifold in a high-dimensional measurement space. This presents both theoretical and computational challenges, which are the focus of this work.

Specifically, we will first describe a new Diffusion EMD that is an $L^1$ distance between density estimates computed using multiple scales of diffusion kernels over a graph. Using theory on the Hölder-Lipschitz dual norm on continuous manifolds (Leeb & Coifman, 2016), we show that as the number of samples increases, Diffusion EMD is equivalent to the Wasserstein distance on the manifold. This formulation reduces the computational complexity of computing K-nearest Wasserstein-neighbors between $m$ distributions over $n$ points from $O(m^2n^3)$ for the exact computation to $\tilde{O}(mn)$ with reasonable assumptions on the data. Finally, we will show how this can be applied to embed large sets of distributions that arise from a common graph, for instance single cell datasets collected on large patient cohorts.

Our contributions include: 1. A new method for computing EMD for distributions over graphs called Diffusion EMD. 2. Theoretical analysis of the relationship between Diffusion EMD and the snowflake of a standard EMD. 3. Fast algorithms for approximating Diffusion EMD. 4. Demonstration of the differentiability of this framework. 5. Application of Diffusion EMD to embedding massive multi-sample biomedical datasets.

## 2. Preliminaries

We now briefly review optimal transport definitions and classic results from diffusion geometry.

**Notation.** We say that two elements $A$ and $B$ are equivalent if there exist $c, C > 0$ such that $cA \leq B \leq CA$, and we denote $A \simeq B$. The definition of $A$ and $B$ will be clear depending on the context.

**Optimal Transport.** Let $\mu, \nu$ be two probability distributions on a measurable space $\Omega$ with metric $d(\cdot, \cdot)$, $\Pi(\mu, \nu)$ be the set of joint probability distributions $\pi$ on the space $\Omega \times \Omega$, where for any subset $\omega \subset \Omega$, $\pi(\omega \times \Omega) = \mu(\omega)$ and $\pi(\Omega \times \omega) = \nu(\omega)$. The 1-Wasserstein distance $W_d$ also known as the earth mover's distance (EMD) is defined as:

$$W_d(\mu, \nu) := \inf_{\pi \in \Pi(\mu, \nu)} \int_{\Omega \times \Omega} d(x, y) \pi(dx, dy). \quad (1)$$

When $\mu, \nu$ are discrete distributions with $n$ points, then Eq. 1 is computable in $O(n^3)$ with a network-flow based algorithm (Peyré & Cuturi, 2019).

Let $\| \cdot \|_{L_d}$ denote the Lipschitz norm w.r.t. $d$, then the dual of Eq. 1 is:

$$W_d(\mu, \nu) = \sup_{\|f\|_{L_d} \leq 1} \int_{\Omega} f(x)\mu(dx) - \int_{\Omega} f(y)\nu(dy). \quad (2)$$

This formulation is known as the Kantorovich dual with $f$ as the witness function. Since it is in general difficult to optimize over the entire space of 1-Lipschitz functions, many works optimize the cost over a modified family of functions such as functions parameterized by clipped neural networks (Arjovsky et al., 2017), functions defined over trees (Le et al., 2019), or functions defined over Haar wavelet bases (Gavish et al., 2010).

**Data Diffusion Geometry** Let $(\mathcal{M}, d_{\mathcal{M}})$ be a connected Riemannian manifold, we can assign to $\mathcal{M}$ sigma-algebras and look at $\mathcal{M}$ as a "metric measure space". We denote by $\Delta$ the Laplace-Beltrami operator on $\mathcal{M}$. For all $x, y \in \mathcal{M}$ let $h_t(x, y)$ be the heat kernel, which is the minimal solution of the heat equation:

$$\left( \frac{\partial}{\partial t} - \Delta_x \right) h_t = 0, \quad (3)$$

with initial condition $\lim_{t \to 0} h_t(x, y) = \delta_y(x)$, where $x \mapsto \delta_y(x)$ is the Dirac function centered at $y$, and $\Delta_x$ is taken with respect to the $x$ argument of $h_t$. Note that as shown in Grigor'yan et al. (2014), the heat kernel captures the local intrinsic geometry of $\mathcal{M}$ in the sense that as $t \to 0$, $\log h_t(x, y) \simeq -d_{\mathcal{M}}^2(x, y)/4t$.

Here, in Sec. 4.2 (Theorem 1) we discuss another topological equivalent of the geodesic distance with a diffusion distance derived from the heat operator $\boldsymbol{H}_t := e^{-t\Delta}$ that characterizes the solutions of the heat equation (Eq. 3), and is related to the heat kernel via $\boldsymbol{H}_t f = \int h_t(\cdot, y)f(y)dy$ (see Lafferty & Lebanon, 2005; Coifman & Lafon, 2006; Grigor'yan et al., 2014, for further details).

It is often useful (particularly in high dimensional data) to consider data as sampled from a lower dimensional manifold embedding in the ambient dimension. This manifold can be characterized by its local structure and in particular, how heat propagates along it. Coifman & Lafon (2006) showed how to build such a propagation structure over discrete data by first building a graph with affinities

$$(\boldsymbol{K}_\epsilon)_{ij} := e^{-\|x_i - x_j\|_2^2/\epsilon} \quad (4)$$

then considering the density normalized operator $\boldsymbol{M}_\epsilon := \boldsymbol{Q}^{-1}\boldsymbol{K}_\epsilon\boldsymbol{Q}^{-1}$, where $\boldsymbol{Q}_{ii} := \sum_j (\boldsymbol{K}_\epsilon)_{ij}$. Lastly, a Markov diffusion operator is defined by

$$\boldsymbol{P}_\epsilon := \boldsymbol{D}^{-1}\boldsymbol{M}_\epsilon, \text{ where } \boldsymbol{D}_{ii} := \sum_j (\boldsymbol{M}_\epsilon)_{ij}. \quad (5)$$

Both $\boldsymbol{D}$ and $\boldsymbol{Q}$ are diagonal matrices. By the law of large numbers, the operator $\boldsymbol{P}_\epsilon$ admits a natural continuous equivalent $\tilde{\boldsymbol{P}}_\epsilon$, i.e., for $n$ i.i.d. points, the sums modulo $n$ converge to the integrals. Moreover, in Coifman & Lafon (2006, Prop. 3) it is shown that $\lim_{\epsilon \to 0} \tilde{\boldsymbol{P}}_\epsilon^{t/\epsilon} = e^{-t\Delta} = \boldsymbol{H}_t$. In conclusion, the operator $\boldsymbol{P}_\epsilon$ converges to $\tilde{\boldsymbol{P}}_\epsilon$ as the sample size increases and $\tilde{\boldsymbol{P}}_\epsilon$ provide an approximation of the Heat kernel on the manifold. Henceforth, we drop the subscript of $\boldsymbol{P}_\epsilon$ to lighten the notation, further we will use the notation $\boldsymbol{P}_\epsilon$ for the operator as well as for the matrix (it will be clear in the context).

## 3. EMD through the $L^1$ metric between multiscale density estimates

The method of efficiently approximating EMD that we will consider here is the approximation of EMD through density estimates at multiple scales. Previous work has considered densities using multiscale histograms over images (Indyk & Thaper, 2003), wavelets over images and trees (Shirdhonkar & Jacobs, 2008; Gavish et al., 2010) and densities over hierarchical clusters (Le et al., 2019). Diffusion EMD also uses a hierarchical set of bins at multiple scales but with smooth bins determined by the heat kernel, which allows us to show equivalence to EMD with a ground distance of the manifold geodesic. These methods are part of a family that we call *multiscale earth mover's distances* that first compute a set of multiscale density estimates or histograms where $L^1$ differences between density estimates realizes an effective *witness function* and have (varying) equivalence to the Wasserstein distance between the distributions. This class of multiscale EMDs are particularly useful in computing embeddings of distributions where the Wasserstein distance is the ground distance, as they are amenable to fast nearest neighbor queries. We explore this application further in Sec. 4.6 and these related methods in Sec. B.1 of the Appendix.

## 4. Diffusion Earth Mover's Distance

We now present the Diffusion EMD, a new Earth Mover's distance based on multiscale diffusion kernels as depicted in Fig. 1. We first show how to model multiple datasets as distributions on a common data graph and perform multiscale density estimates on this graph.

### 4.1. Data Graphs and Density Estimates on Graphs

Let $\mathcal{X} = \{X_1, X_2, \ldots, X_m\}$, $\cup_{j=1}^m X_j \subseteq \mathcal{M} \subset \mathbb{R}^d$, be a collection of datasets with $n_i = |X_i|$ and $n = \sum_i n_i$. Assume that the $X_i$'s are independently sampled from a common underlying manifold $(\mathcal{M}, d_{\mathcal{M}})$ which is a Riemannian closed manifold (compact and without boundary) immersed in a (high dimensional) ambient space $\mathbb{R}^d$, with

geodesic distance $d_{\mathcal{M}}$. Further, assume that while the underlying manifold is common, each dataset is sampled from a different distribution over it, as discussed below. Such collections of datasets arise from several related samples of data, for instance single cell data collected on a cohort of patients with a similar condition.

Here, we consider the datasets in $\mathcal{X}$ as representing distributions over the common data manifold, which we represent in the finite setting as a common data graph $G_{\mathcal{X}} = (V, E, w)$ with $V = \cup_{j=1}^m X_j$ and edge weights determined by the Gaussian kernel (see Eq. 4), where we identify edge existence with nonzero weights. Then, we associate each $X_i$ with a density measure $\boldsymbol{\mu}_i^{(t)} : V \to [0, 1]$, over the entire data graph. To compute such measures, we first create indicator vectors for the individual datasets on it, let $\mathbf{1}_{X_i} \in \{0, 1\}^n$ be a vector where for each $v \in V$, $\mathbf{1}_{X_i}(v) = 1$ if and only if $v \in X_i$. We then derive a kernel density estimate by applying the diffusion operator constructed via Eq. 5 over the graph $G$ to these indicator functions to get scale-dependent estimators

$$\boldsymbol{\mu}_i^{(t)} := \frac{1}{n_i} \boldsymbol{P}^t \mathbf{1}_{X_i}, \qquad (6)$$

where the scale $t$ is the diffusion time, which can be considered as a meta-parameter (e.g., as used in Burkhardt et al., 2021) but can also be leveraged in multiscale estimation of distances between distributions as discussed here. Indeed, as shown in Burkhardt et al. (2021), at an appropriately tuned single scale, this density construction yields a discrete version of kernel density estimation.

### 4.2. Diffusion Earth Mover's Distance Formulation

We define the Diffusion Earth Mover's Distance between two datasets $X_i, X_j \in \mathcal{X}$ as

$$W_{\alpha, K}(X_i, X_j) := \sum_{k=0}^{K} \|T_{\alpha,k}(X_i) - T_{\alpha,k}(X_j)\|_1 \quad (7)$$

where $0 < \alpha < 1/2$ is a meta-parameter used to balance long- and short-range distances, which in practice is set close to 1/2, $K$ is the maximum scale considered here, and

$$T_{\alpha,k}(X_i) := \begin{cases} 2^{-(K-k-1)\alpha}(\boldsymbol{\mu}_i^{(2^{k+1})} - \boldsymbol{\mu}_i^{(2^k)}) & k < K \\ \boldsymbol{\mu}_i^{(2^K)} & k = K \end{cases}$$

$$(8)$$

Further, to set $K$, we note that if the Markov process governed by $\boldsymbol{P}$ converges (i.e., to its stationary steady state) in polynomial time w.r.t. $|V|$, then one can ensure that beyond $K = O(\log |V|)$, all density estimates would be essentially indistinguishable as shown by the following lemma, whose proof appears in the Appendix:
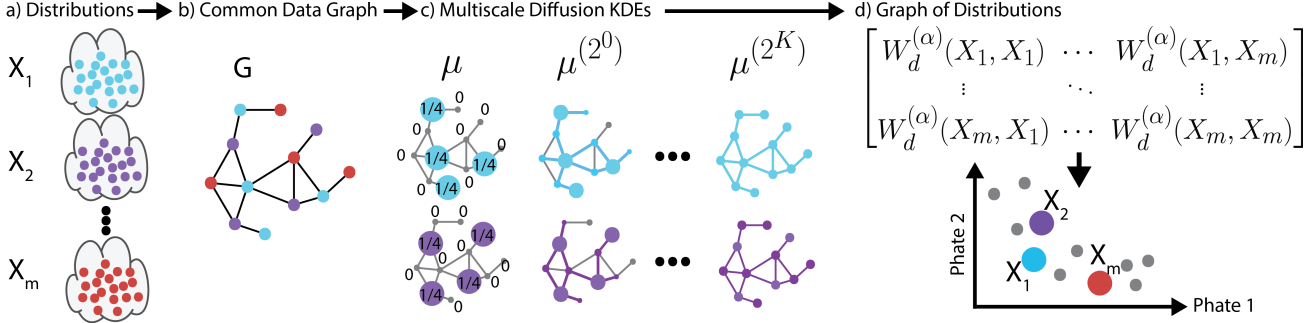
Figure 1. Diffusion EMD first embeds datasets into a common data graph $G$, then takes multiscale diffusion KDEs for each of the datasets. These multiscale KDEs are then used to compute the Diffusion Earth Mover's Distance between the datasets that can be used in turn to create graphs and embeddings (PHATE (Moon et al., 2018) shown here) of the datasets.

**Lemma 1.** *There exists a* $K = O(\log|V|)$ *such that* $\boldsymbol{\mu}_i^{(2^K)} \simeq \boldsymbol{P}^{2^K} \mathbf{1}_{X_i} \simeq \phi_0$ *for every* $i = 1, \ldots, n$, *where* $\phi_0$ *is the trivial eigenvector of* $\boldsymbol{P}$ *associated with the eigenvalue* $\lambda_0 = 1$.

To compute the Diffusion EMD $W_{\alpha,K}$ in Eq. 7 involves first calculating diffusions of the datasets $\boldsymbol{\mu}$, second calculating differences and weighting them in relation to their scale, this results in a vector per distribution of length $O(nK)$, and finally computing the $L^1$ norm between them. The most computationally intensive step, computing the diffusions, is discussed in further detail in Sec. 4.4.

### 4.3. Theoretical Relation to EMD on Manifolds

We now provide a theoretical justification of the Diffusion EMD defined via Eq. 7 by following the relation established in Leeb & Coifman (2016) between heat kernels and the EMD on manifolds. Leeb & Coifman (2016) define the following ground distance for EMD over $\mathcal{M}$ by leveraging the geometric information gathered from the $L^1$ distances between kernels at different scales.

**Definition 1.** *The diffusion ground distance between* $x, y \in \mathcal{M}$ *is defined as*

$$D_\alpha(x,y) := \sum_{k \geq 0} 2^{-k\alpha} \|h_{2^{-k}}(x,\cdot) - h_{2^{-k}}(y,\cdot)\|_1,$$

*for* $\alpha \in (0, 1/2)$, *the scale parameter* $K \geq 0$ *and* $h_t(\cdot, \cdot)$ *the heat kernel on* $\mathcal{M}$.

Note that $D_\alpha(\cdot, \cdot)$ is similar to the diffusion distance defined in Coifman & Lafon (2006), which was based on $L^2$ notions rather than $L^1$ here. Further, the following result from Leeb & Coifman (2016, see Sec. 3.3) shows that the distance $D_\alpha(\cdot, \cdot)$ is closely related to the intrinsic geodesic one $d_\mathcal{M}(\cdot, \cdot)$.

**Theorem 1.** *Let* $(\mathcal{M}, d_\mathcal{M})$ *be a closed manifold with geodesic distance* $d_\mathcal{M}$, *and let* $\alpha \in (0, 1/2)$. *The metric* $D_\alpha(\cdot, \cdot)$ *defined via Def. 1 is equivalent to* $d_\mathcal{M}(\cdot, \cdot)^{2\alpha}$.

The previous theorem justifies why in practice we let $\alpha$ close to $1/2$, because we want the snowflake distance $d_\mathcal{M}^{2\alpha}(\cdot, \cdot)$ to approximate the geodesic distance of $\mathcal{M}$. The notion of equivalence established by this result is such that $D_\alpha(x, \cdot) \simeq d(x, \cdot)^{2\alpha}$. It is easy to verify that two equivalent metrics induce the same topology. We note that while here we only consider the Heat kernel, a similar result holds (see Theorem 4 in the Appendix) for a more general family of kernels, as long as they satisfy certain regularity conditions.

For a family of operators $(\boldsymbol{A}_t)_{t \in \mathbb{R}^+}$ we define the following metric on distributions; let $\mu$ and $\nu$ be two distributions:

$$\widehat{W}_{\boldsymbol{A}_t}(\mu, \nu) = \|\boldsymbol{A}_1(\mu - \nu)\|_1 \tag{9}$$
$$+ \sum_{k \geq 0} 2^{-k\alpha} \|(\boldsymbol{A}_{2^{-(k+1)}} - \boldsymbol{A}_{2^{-k}})(\mu - \nu)\|_1.$$

The following result shows that applying this metric for the family of operators $\boldsymbol{H}_t$ to get $\widehat{W}_{\boldsymbol{H}_t}$ yields an equivalent of the EMD with respect to the diffusion ground distance $D_\alpha(\cdot, \cdot)$.

**Theorem 2.** *The EMD between two distributions* $\mu, \nu$ *on a closed Riemannian manifold* $(\mathcal{M}, d_\mathcal{M})$ *w.r.t. the diffusion ground distance* $D_\alpha(\cdot, \cdot)$, *defined via Def. 1, given by*

$$W_{D_\alpha}(\mu, \nu) = \inf_{\pi \in \Pi(\mu, \nu)} \int_{\mathcal{M} \times \mathcal{M}} D_\alpha(x, y) \pi(dx, dy), \tag{10}$$

*is equivalent to* $\widehat{W}_{\boldsymbol{H}_t}$. *That is* $W_{D_\alpha} \simeq \widehat{W}_{\boldsymbol{H}_t}$, *where* $\boldsymbol{H}_t$ *is the Heat operator on* $\mathcal{M}$.

*Proof.* In Proposition 15 of Leeb & Coifman (2016), it is shown that $\mathcal{M}$ is separable w.r.t. $D_\alpha(\cdot, \cdot)$, hence we can use the Kantorovich-Rubinstein theorem. We let $\Lambda_\alpha$, the space of functions that are Lipschitz w.r.t. $D_\alpha(\cdot, \cdot)$ and $\|\cdot\|_{\Lambda_\alpha^*}$, the norm of its dual space $\Lambda_\alpha^*$. The norm is defined by

$$\|T\|_{\Lambda_\alpha^*} := \sup_{\|f\|_{\Lambda_\alpha} \leq 1} \int_\mathcal{M} f \, dT.$$

In Theorem 4 of Leeb & Coifman (2016), it is shown that both $W_{D_\alpha}$ and $\widehat{W}_{H_t}$ are equivalent to the norm $\|\cdot\|_{\Lambda_\alpha^*}$. $\qquad\square$

We consider the family of operators $(P_\epsilon^{t/\epsilon})_{t\in\mathbb{R}^+}$, which is related to the continuous equivalent of the stochastic matrix defined in Eq. 5. In practice, we use this family of operators to approximate the heat operator $H_t$. Indeed, when we take a small value of $\epsilon$, as discussed in Sec. 2, we have from Coifman & Lafon (2006) that this is a valid approximation.

**Corollary 2.1.** *Let $P_\epsilon$ be the continuous equivalent of the stochastic matrix in Eq. 5. For $\epsilon$ small enough, we have:*

$$\widehat{W}_{P_\epsilon^{t/\epsilon}} \simeq W_{D_\alpha}. \tag{11}$$

Eq. 11 motivates our use of Eq. 7 to compute the Diffusion EMD. The idea is to take only the first $K$ terms in the infinite sum $\widehat{W}_{P_\epsilon^{t/\epsilon}}$ and then choosing $\epsilon := 2^{-K}$ would give us exactly Eq. 7. We remark that the summation order of Eq. 7 is inverted compared to Eq. 11, but in both cases the largest scale has the largest weight. Finally, we state one last theorem that brings our distance closer to the Wasserstein w.r.t. $d_{\mathcal{M}}(\cdot,\cdot)$; we refer the reader to the Appendix for its proof.

**Theorem 3.** *Let $\alpha \in (0, 1/2)$ and $(\mathcal{M}, d_{\mathcal{M}})$ be a closed manifold with geodesic $d_{\mathcal{M}}$. The Wasserstein distance w.r.t. the diffusion ground distance $D_\alpha(\cdot, \cdot)$ is equivalent to the Wasserstein distance w.r.t the snowflake distance $d_{\mathcal{M}}(\cdot, \cdot)^{2\alpha}$ on $\mathcal{M}$, that is $W_{D_\alpha} \simeq W_{d_{\mathcal{M}}^{2\alpha}}$.*

**Corollary 3.1.** *For each $1 \leq i, j \leq m$, let $X_i, X_j \in \mathcal{X}$ be two datasets with size $n_i$ and $n_j$ respectively, and let $\mu_i$ and $\mu_j$ be the continuous distributions corresponding to the ones of $X_i$ and $X_j$, let $K$ be the largest scale and put $N = \min(K, n_i, n_j)$. Then, for sufficiently big $N \to \infty$ (implying sufficiently small $\epsilon = 2^{-K} \to 0$):*

$$W_{\alpha,K}(X_i, X_j) \simeq W_{d_{\mathcal{M}}^{2\alpha}}(\mu_i, \mu_j), \tag{12}$$

*for all $\alpha \in (0, 1/2)$.*

In fact we can summarize our chain of thought as follows

$$W_{\alpha,K}(X_i, X_j) \overset{(a)}{\simeq} \widehat{W}_{P_\epsilon^{t/\epsilon}}(\mu_i, \mu_j) \overset{(b)}{\simeq} \widehat{W}_{H_t}(\mu_i, \mu_j)$$
$$\overset{(c)}{\simeq} W_{D_\alpha}(\mu_i, \mu_j) \overset{(d)}{\simeq} W_{d_{\mathcal{M}}^{2\alpha}}(\mu_i, \mu_j),$$

where the approximation *(a)* is due to the fact that the discrete distributions on $X_i$ and $X_j$ converge respectively to $\mu_i$ and $\mu_j$ when $\min(n_i, n_j) \to \infty$. Further, $W_{\alpha,K}(X_i, X_j)$ approximate the infinite series $\widehat{W}_{P_\epsilon^{t/\epsilon}}(\mu_i, \mu_j)$ as in Eq. 9 when $K \to \infty$, note also that we take $\epsilon = 2^{-K}$ so that the largest scale in Eq. 7 is exactly $2^K$. The approximation in *(b)* comes from the approximation of the heat operator as in Coifman & Lafon (2006), *(c)* comes from Theorem 2 and *(d)* comes from Theorem 1.

**Algorithm 1** Chebyshev embedding

**Input:** $n \times n$ graph kernel $K$, $n \times m$ distributions $\mu$, maximum scale $K$, and snowflake constant $\alpha$.
**Output:** $m \times (K+1)n$ distribution embeddings $b$
$Q \leftarrow Diag(\sum_i K_{ij})$
$K^{norm} \leftarrow Q^{-1} K Q^{-1}$
$D \leftarrow Diag(\sum_i K_{ij}^{norm})$
$M \leftarrow D^{-1/2} K^{norm} D^{-1/2}$
$U \Sigma U^T = M$; $\quad U$ orthogonal, $\Sigma$ Diagonal
$\mu^{(2^0)} \leftarrow P\mu \leftarrow D^{-1/2} M D^{1/2} \mu$
**for** $k = 1$ **to** $K$ **do**
$\quad \mu^{(2^k)} \leftarrow P^{2^k}\mu \leftarrow D^{-1/2} U (\Sigma)^{2^k} U^T D^{1/2} \mu$
$\quad b_{k-1} \leftarrow 2^{(K-k-1)\alpha}(\mu^{(2^k)} - \mu^{(2^{k-1})})$
**end for**
$b_K \leftarrow \mu^{(2^K)}$
$b \leftarrow [b_0, b_1, \ldots, b_K]$

## 4.4. Efficient Computation of Dyadic Scales of the Diffusion Operator

The most computationally intensive step of Diffusion EMD requires computing dyadic scales of the diffusion operator $P$ times $\mu$ to estimate the density of $\mu^{(t)}$ at multiple scales which we will call $b$. After this embedding, Diffusion EMD between two embeddings $b_i, b_j$ is computed as $|b_i - b_j|$, i.e. the $L^1$ norm of the difference. Computing the embedding $b$ naively by first powering $P$ then right multiplying $\mu$, may take up to $2^K$ matrix multiplications which is infeasible for even moderately sized graphs. We assume two properties of $P$ that makes this computation efficient in practice. First, that $P$ is sparse with order $\tilde{O}(n)$ non-zero entries. This applies when thresholding $K_\epsilon$ or when using a $K$-nearest neighbors graph to approximate the manifold (van der Maaten & Hinton, 2008; Moon et al., 2019). Second, that $P^t$ is low rank for large powers of $t$.

While there are many ways to approximate dyadic scales of $P$, we choose from two methods depending on the number of distributions $m$ compared to the number of points in the graph $n$. When $m \ll n$, we use a method based on Chebyshev approximation of polynomials of the eigenspectrum of $P$ as shown in Alg. 1. This method is efficient for sparse $P$ and a small number of distributions (Shuman et al., 2011). For more detail on the error incurred by using Chebyshev polynomials we refer the reader to Trefethen (2013, Chap. 3). In practice, this requires for the approximating polynomial of $J$ terms, computation of $mJ$ (sparse) matrix vector multiplications for a worst case time complexity of $O(Jmn^3)$, but in practice is $\tilde{O}(Jmn)$ where $J$ is a small constant (see Fig. 5(e)). However, while asymptotically efficient, when $m \gg n$ in practice this can be inefficient as it requires many multiplications of the form $P\mu$.

In the case where $m \gg n$, approximating powers of $P$ and

applying these to the $n \times m$ collection of distributions $\boldsymbol{\mu}$ once is faster. This method is also useful when the full set of distributions is not known and can be applied to new distributions one at a time in a data streaming model. A naive approach for computing $\boldsymbol{P}^{2^K}$ would require $K$ dense $n \times n$ matrix multiplications. However, as noted in Coifman & Maggioni (2006), for higher powers of $\boldsymbol{P}$, we can use a much smaller basis. We use an algorithm based on interpolative decomposition (Liberty et al., 2007; Bermanis et al., 2013) to reduce the size of the basis, and subsequently the computation time, at higher scales. In this algorithm we first determine the approximate rank of $\boldsymbol{P}^{2^k}$ using an estimate of the density of the eigenspectrum of $\boldsymbol{P}$ as in Dong et al. (2019). We then alternate steps of downsampling the basis to the specified rank of $\boldsymbol{P}^{2^k}$ with (randomized) interpolative decomposition with steps of powering $\boldsymbol{P}$ on these bases. Informally, the interpolative decomposition selects a representative set of points that approximate the basis well. In the worst case this algorithm can take $\tilde{O}(mn^3)$ time to compute the diffusion density estimates, nevertheless with sparse $\boldsymbol{P}$ with a rapidly decaying spectrum, this algorithm is $\tilde{O}(mn)$ in practice. For more details see Alg. 4 and Sec. C of the Appendix.

## 4.5. Subsampling Density Estimates

The density estimates created for each distribution are both large and redundant with each distribution represented by a vector of $(K + 1) \times n$ densities. However, as noted in the previous section, $\boldsymbol{P}^{2^k}$ can be represented on a smaller basis, especially for larger scales. Intuitively, the long time diffusions of nodes that are close to each other are extremely similar. Interpolative decomposition (Liberty et al., 2007; Bermanis et al., 2013) allows us to pick a set of points to center our diffusions kernels such that they approximately cover the graph up to some threshold on the rank. In contrast, in other multiscale EMD methods the bin centers or clusters are determined randomly, making it difficult to select the number of centers necessary. Furthermore, the relative number of centers at every scale is fixed, for example, Quadtree or Haar wavelet based methods (Indyk & Thaper, 2003; Gavish et al., 2010) use $2^{d^k}$ centers at every scale, and a clustering based method (Le et al., 2019) selects $C^k$ clusters at every scale for some constant $C$. Conversely, in Diffusion EMD, by analyzing the behavior of $\boldsymbol{P}^{2^k}$, we intelligently select the number of centers needed at each scale based on the approximate rank of $\boldsymbol{P}^{2^k}$ up to some tolerance at each scale. This does away with the necessity of a fixed ratio of bins at every scale, allowing adaptation depending on the structure of the manifold and can drastically reduce the size representations (see Fig. 5(c)). For the Chebyshev polynomials method, this subsampling is done post computation of diffusion scales, and for the method based on approximating

$\boldsymbol{P}^{2^k}$ directly the subsampling happens during computation. To this point, we have described a method to embed distributions on a graph into a set of density estimates whose size depends on the data, and the spectrum decay of $\boldsymbol{P}$. We will now explore how to use these estimates for exploring the Diffusion EMD metric between distributions.

## 4.6. Diffusion EMD Based Embeddings of Samples

Our main motivation for a fast EMD computed on related datasets is to examine the space of the samples or datasets themselves, i.e., the higher level manifold of distributions. In terms of the clinical data, on which we show this method, this would be the relationship between patients themselves, as determined by the EMD between their respective single-cell peripheral blood datasets. Essentially, we create a kernel matrix $\boldsymbol{K}_{\mathcal{X}}$ and diffusion operator $\boldsymbol{P}_{\mathcal{X}}$ between datasets where the samples are nodes on the associated graph. This diffusion operator $\boldsymbol{P}_{\mathcal{X}}$ can be embedded using diffusion maps (Coifman & Lafon, 2006) or visualized with a method like PHATE (Moon et al., 2018) that collects the information into two dimensions as shown in Sec. 5. We note this higher level graph can be a sparse KNN graph, particularly given that a diffusion operator on the graph can allow for global connections to be reformed via $t$-step path probabilities.

Multiscale formulations of EMD as in Eq. 8 are especially effective when searching for nearest neighbor distributions under the Wasserstein metric (Indyk & Thaper, 2003; Backurs et al., 2020) as this distance forms a normed space, i.e., a space where the metric is induced by the $L_1$ norm of the distribution vectors and their differences. Data structures such as kd-trees, ball trees, locality sensitive hashing, are able to take advantage of such normed spaces for sub-linear neighbor queries. This is in contrast to network-flow or Sinkhorn type approximations that require a scan through all datapoints for each nearest neighbor query as this metric is not derived from a norm.

## 4.7. Gradients of the Earth Mover's Distance

One of the hindrances in the use of optimal transport-based distances has been the fact that it cannot be easily incorporated into deep learning frameworks. Gradients with respect to the EMD are usually found using a trained Lipschitz discriminator network as in Wasserstein-GANs (Arjovsky et al., 2017), which requires unstable adversarial training, or by taking derivatives through a small number of iterations of the Sinkhorn algorithm (Frogner et al., 2015; Bonneel et al., 2016; Genevay et al., 2018; Liu et al., 2020), which scales with $O(n^2)$ in the number of points. Tree based methods that are linear in the number of points do not admit useful gradients due to their hard binning over space, giving a gradient of zero norm almost everywhere.

We note that, given the data diffusion operator, the com-

putation of Diffusion EMD is differentiable and, unlike Tree-based EMD, has smooth bins and therefore a non-zero gradient norm near the data (as visible in Fig. 2). Further, computation of the gradient only requires powers of the diffusion operator multiplied by the indicator vector describing the distribution on the graph. In fact, as mentioned in the supplementary material (Sec. C.1) for each $v \in V$, the gradient of the Diffusion EMD $\partial W_{\alpha, K}(X_i, X_j)/\partial v$ depends mainly on the gradients $\partial P_\epsilon^{2^k}/\partial v$ for $0 \leq K$ which can be expressed in terms of the gradient of the Gaussian kernel $\partial K_\epsilon/\partial v$. This last quantity is easy to compute. In Sec. C.1 of the Appendix, we give an exact process on computing the gradient of the Diffusion EMD.

## 5. Results

In this section, we first evaluate the Diffusion EMD on two manifolds where the ground truth EMD with a geodesic ground distance is known, a swiss roll dataset and spherical MNIST (Cohen et al., 2017). On these datasets where we have access to the ground truth geodesic distance we show that Diffusion EMD is both faster and closer to the ground truth than comparable methods. Then, we show an application to a large single cell dataset of COVID-19 patients where the underlying metric between cells is thought to be a manifold (Moon et al., 2018; Kuchroo et al., 2020). We show that the manifold of patients based on Diffusion EMD by capturing the graph structure, better captures the disease state of the patients.

**Experimental Setup.** We consider four baseline methods for approximating EMD: QuadTree($D$) (Backurs et al., 2020) which partitions the dataspace in half in each dimension up to some specified depth $D$, ClusterTree($C, D$) (Le et al., 2019) which recursively clusters the data with $C$ clusters up to depth $D$ using the distance between clusters to weight the tree edges, and the convolutional Sinkhorn distance (Solomon et al., 2015) with the same graph as used in Diffusion EMD. QuadTree and ClusterTree are fast to compute. However, because they operate in the ambient space, they do not represent the geodesic distances on the manifold in an accurate way. The convolutional Sinkhorn method represents the manifold well but is significantly slower even when using a single iteration. For more details on related work and the experimental setup see Sections B and D of the Appendix respectively.

**1D data.** We first illustrate the advantages of using Diffusion EMD over tree-based methods (ClusterTree and QuadTree) on a line graph with 500 points spaced in the interval $[0, 1]$. In Fig. 2 we depict the Wasserrstein distance between an indicator function at each of these 500 points ($1_x$) and an indicator function at $x = 0.5$. In the case of indicator functions the Wasserstein distance is exactly the
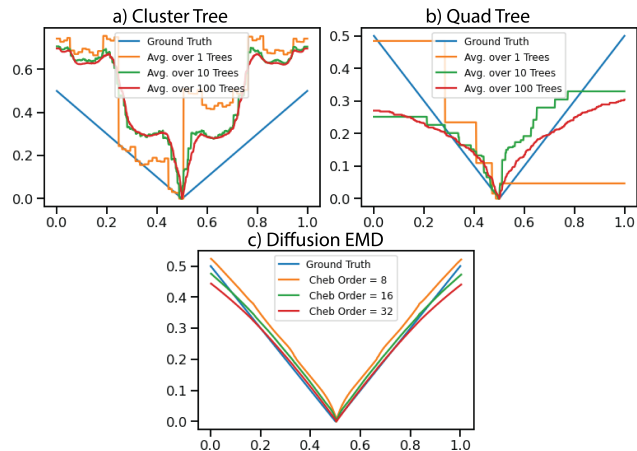


*Figure 2.* Wasserstein distance of indicator distributions $1_x, x \in [0, 1]$ from $1_{0.5}$ computed using linear EMD methods $L^2$ distance: (a) ClusterTree (b) QuadTree and (c) Diffusion EMD.

ground distance. We show three approximations of each method, varying the number of trees and the Chebyshev polynomial order respectively. It is clear that Diffusion EMD achieves a much better approximation of the ground truth primarily due to its use of smooth bins.
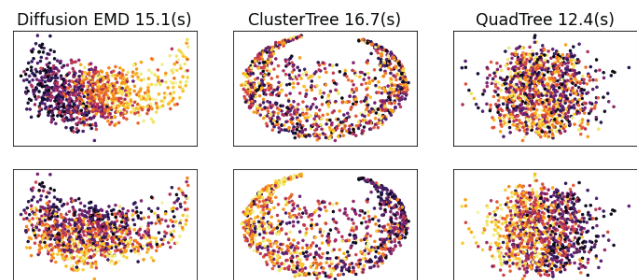


*Figure 3.* Swiss roll dataset embeddings of $m = 1000$ distributions with $n = 10,000$ total points rotated into 10D colored by ground truth 2D sheet axes. Diffusion EMD recreates the manifold better in similar time.

**Swiss roll data.** The next application we explore is to a dataset where we have distributions on a manifold for which the geodesic distance is easily computable. In this way we can compare to the ground truth EMD between distributions. We generate $m = 100$ Gaussians on the swiss roll with 100 points each for a total of $n = 10,000$ points on the graph. We compare each method on two metrics, the 10-nearest neighbor accuracy measured (P@10) where a P@10 of 1 means that the 10-nearest neighbors are the same as the ground truth. We compare the rankings of nearest neighbors over the entire dataset using the Spearman-$\rho$ correlation coefficient, which measures the similarity of nearest neighbor rankings. This coefficient ranges between -1 for inversely ranked lists and 1 for the same ranks. This measures rankings over the entire dataset equally rather
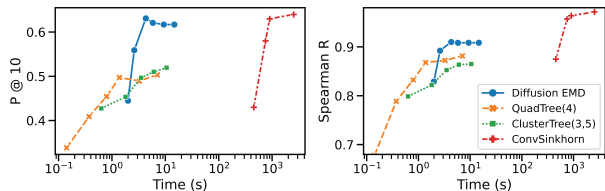
*Figure 4.* Accuracy of methods measured via P@10 (left) and Spearman coefficient (right), against their (log scaled) computation time in seconds on the swiss roll dataset. Variations of methods are over Chebyshev approximation order for Diffusion EMD, # of trees for tree methods, and number of iterations for conv. Sinkhorn. Diffusion EMD is more accurate than tree methods and orders of magnitude faster than conv. Sinkhorn even with a single iteration.
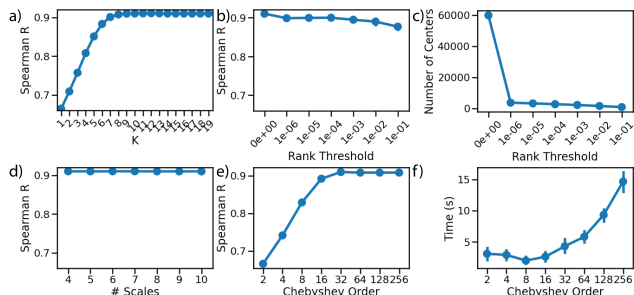


*Figure 5.* Ablation study of major parameters for Chebyshev polynomial approximation on the swiss roll dataset. Mean and std. over 10 runs over (a) values of the maximum scale $K$, (b) the rank threshold in interpolative decomposition, (c) the total number of centers in the $L^1$ representation which drops with decomposition, (d-e) performance against the # of scales, and the order of the polynomial, both are very stable after a certain point, and (f) time vs. the Chebyshev order.

than only considering the nearest neighbors. Visually, we show embeddings of the swiss roll in Fig. 3, where the 2D manifold between distributions is best captured by Diffusion EMD given a similar amount of time.

In Fig. 4, we investigate the time vs. accuracy tradeoff of a number of fast EMD methods on the swiss roll. We compare against the ground truth EMD which is calculated with the exact EMD on the "unrolled" swiss roll in 2D. We find that Diffusion EMD is more accurate than tree methods for a given amount of time and is much faster than the convolutional Sinkhorn method and only slightly less accurate. To generate multiple models for each dataset we vary the number of trees for tree methods, the Chebyshev order for Diffusion EMD, and the number of iterations for convolutional Sinkhorn. We search over and fix other parameters using a grid search as detailed in Sec. D of the Appendix.

In Fig. 5, we vary parameters of the Chebyshev polynomial algorithm of Diffusion EMD. Regarding performance, we find Diffusion EMD is stable to the number of scales chosen after a certain minimum maximum scale $K$, Chebyshev

polynomial order, and the number of scales used. By performing interpolative decomposition with a specified rank threshold on $P^{2^k}$ we can substantially reduce the embedding size at a small cost to performance Fig. 5(b,c).

*Table 1.* Classification accuracy, P@10, Spearman $\rho$ and runtime (in minutes) on 70,000 distributions from Spherical MNIST.

|  | ACCURACY | P@10 | SPEARMAN $\rho$ | TIME |
|---|---|---|---|---|
| DIFF. EMD | **95.94** | **0.611** | **0.673** | 34M |
| CLUSTER | 91.91 | 0.393 | 0.484 | 30M |
| QUAD | 79.56 | 0.294 | 0.335 | **16M** |

**Spherical MNIST.** Next, we use the Spherical MNIST dataset to demonstrate the efficacy of the interpolative decomposition based approximation to Diffusion EMD, as here $m \gg n$. Each image is treated as a distribution (of pixel intensities) over the sphere. To evaluate the fidelity of each embedding, we evaluate the 1-NN classification on the embedding vectors in addition to P@10 and Spearman coefficient in Tab. 1. Diffusion EMD creates embeddings that better approximate true EMD over the sphere than tree methods in a similar amount of time, which in this case also gives better classification accuracy.
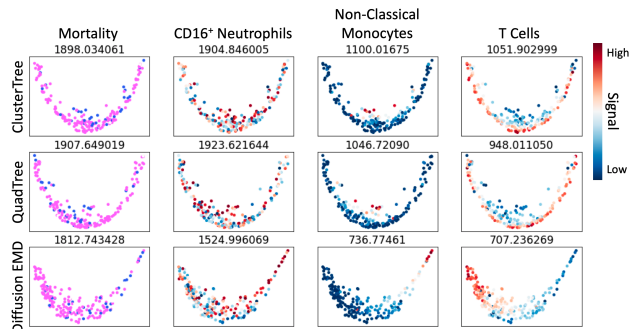


*Figure 6.* Embedding of 210 patients through different manifold constructions. Visualizing patient eventual mortality and cell types predictive of disease outcome on each manifold. Laplacian smoothness reported on each signal for each manifold.

**Single cell COVID-19 patient data.** The COVID-19 pandemic has driven biologists to generate vast amounts of cellular data on hospitalized patients suffering from severe disease. A major question in clinicians minds is determining a priori which patients may be at risk for worse outcomes, including requiring increased ventilatory support and increased risk of mortality. Certain cell types found in the blood, such as $CD16^+$ Neutrophils, T cells and non-classical monocytes, have been associated with and predictive of mortality outcome. Ideally, a manifold of patients would find these cellular populations to occupy a region of high mortality for $CD16^+$ Neutrophils and non-classical monocytes

and low mortality for T cells. In order to construct a manifold of patients suffering from COVID-19, we analyzed 210 blood samples from 168 patients infected with SARS-CoV-2 measured on a myeloid-specific flow cytometry panel, an expanded iteration of a previously published dataset (Lucas et al., 2020). We embedded 22 million cells from these patients into a common combined cell-cell graph with 27,000 nodes as defined in Sec. 4.1. We then computed Diffusion EMD and other methods on these datasets. Diffusion EMD is computed by using indicator vectors for each patient converted to density estimates as in Eq. 6.

On an informative embedding of patients, similar patients, with similar features (such as mortality) would localize on the manifold and thus the important features should be smooth over the manifold. Furthermore, cell types which are correlated with outcome either positively or negatively should also be smooth and either correlated or anticorrelated with outcome. To quantify this, we compute the smoothness with respect to the patient manifold by using a *Laplacian quadratic form* with respect to the 10-NN graph between patients. Convolutional Sinkhorn does not scale to this data, so we compare a patient manifold created with Diffusion EMD to ones created with QuadTree and ClusterTree. Diffusion EMD is able to use the manifold of cells where QuadTree and ClusterTree are built in the ambient space. In Fig. 6 we visualize relevant signals over the patients using PHATE (Moon et al., 2019) overlayed with the quadratic smoothness of the signal over the graph. While the mortality signal appeared enriched in the right branches of both the Diffusion EMD and QuadTree manifolds, it did not localize as well on the ClusterTree manifold. Both CD16$^+$ Neutrophils and non-classical monocytes appeared smoother over the Diffusion EMD manifold than the comparison manifolds. Since both cell types are associated with mortality, it was interesting to see them both enriched in high mortality region of the Diffusion EMD manifold but not the others. Finally, T cells, which are negatively correlated with mortality appeared smoothly enriched in the Diffusion EMD manifold in a region with no mortality. In QuadTree and ClusterTree constructions, T cells appeared enriched throughout the manifold, no localizing smoothly to a region with low mortality. These experiments show that the patient manifold constructed with Diffusion EMD is more informative, smoothly localizing key signals, such as patient outcome and predictive cell types. For a more details see Sec. D of the Appendix.

## 6. Conclusion

In this work we have introduced Diffusion EMD, a multi-scale distance that uses heat kernel diffusions to approximate the earth mover's distance over a data manifold. We showed how Diffusion EMD can efficiently embed many samples on a graph into a manifold of samples in $\tilde{O}(mn)$ time more accurately than similarly efficient methods. This is useful in the biomedical domain as we show how to embed COVID-19 patient samples into a higher level patient manifold that more accurately represents the disease structure between patients. Finally, we also show how to compute gradients with respect to Diffusion EMD, which opens the possibility of gradients of the earth mover's distance that scale linearly with the dataset size.

## References

Arjovsky, M., Chintala, S., and Bottou, L. Wasserstein generative adversarial networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *PMLR*, pp. 214–223, 2017.

Backurs, A., Dong, Y., Indyk, P., Razenshteyn, I., and Wagner, T. Scalable nearest neighbor search for optimal transport. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *PMLR*, pp. 497–506, 2020.

Benamou, J.-D., Carlier, G., Cuturi, M., Nenna, L., and Peyré, G. Iterative Bregman projections for regularized transportation problems. *SIAM Journal on Scientific Computing*, 37(2):A1111–A1138, 2015.

Bermanis, A., Averbuch, A., and Coifman, R. R. Multiscale data sampling and function extension. *Applied and Computational Harmonic Analysis*, 34(1):15–29, 2013.

Bonneel, N., Peyré, G., and Cuturi, M. Wasserstein barycentric coordinates: Histogram regression using optimal transport. *ACM Transactions on Graphics*, 35(4), 2016.

Burkhardt, D. B., Stanley, J. S., Tong, A., Perdigoto, A. L., Gigante, S. A., Herold, K. C., Wolf, G., Giraldez, A. J., van Dijk, D., and Krishnaswamy, S. Quantifying the effect of experimental perturbations at single-cell resolution. *Nature Biotechnology*, 39:619–629, 2021.

Chen, W. S., Zivanovic, N., van Dijk, D., Wolf, G., Bodenmiller, B., and Krishnaswamy, S. Uncovering axes of

variation among single-cell cancer specimens. *Nature Methods*, 17:302–310, 2020.

Cohen, T., Geiger, M., Köhler, J., and Welling, M. Convolutional networks for spherical signals. In *ICML 2017 Workshop on Principled Approaches to Deep Learning*, 2017.

Coifman, R. R. and Lafon, S. Diffusion maps. *Applied and Computational Harmonic Analysis*, 21(1):5–30, 2006.

Coifman, R. R. and Maggioni, M. Diffusion wavelets. *Applied and Computational Harmonic Analysis*, 21(1):53–94, 2006.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems 26*, pp. 2292–2300, 2013.

Demetci, P., Santorella, R., Sandstede, B., Noble, W. S., and Singh, R. Gromov-wasserstein optimal transport to align single-cell multi-omics data. bioRxiv, DOI: 10.1101/2020.04.28.066787, 2020.

Dong, K., Benson, A. R., and Bindel, D. Network density of states. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 1152–1161, 2019.

Frogner, C., Zhang, C., Mobahi, H., Araya, M., and Poggio, T. A. Learning with a wasserstein loss. In *Advances in Neural Information Processing Systems 28 (NeurIPS 2015)*, 2015.

Gavish, M., Nadler, B., and Coifman, R. R. Multiscale wavelets on trees, graphs and high dimensional data: Theory and applications to semi supervised learning. In *Proceedings of the 27th International Conference on Machine Learning (ICML)*, pp. 367–374, 2010.

Genevay, A., Peyre, G., and Cuturi, M. Learning generative models with sinkhorn divergences. In *Proceedings of the 21st International Conference on Artificial Intelligence and Statistics (AISTAT)*, volume 84 of *PMLR*, pp. 1608–1617, 2018.

Grigor'yan, A. and Liu, L. Heat kernel and Lipschitz-Besov spaces. *Forum Mathematicum*, 27(6):3567–3613, 2015.

Grigor'yan, A., Hu, J., and Lau, K.-S. Heat kernels on metric measure spaces. In *Geometry and Analysis of Fractals*, volume 88 of *PROMS*, pp. 147–207, 2014.

Indyk, P. and Thaper, N. Fast image retrieval via embeddings. In *the 3rd International Workshop on Statistical and Computational Theories of Vision*, 2003.

Kuchroo, M., Huang, J., Wong, P., Grenier, J.-C., Shung, D., Tong, A., Lucas, C., Klein, J., Burkhardt, D., Gigante, S., Godavarthi, A., Israelow, B., Mao, T., Oh, J. E., Silva, J., Takahashi, T., Odio, C. D., Casanovas-Massana, A., Fournier, J., the Yale IMPACT Team, Farhadian, S., Dela Cruz, C. S., Ko, A. I., Wilson, F. P., Hussin, J., Wolf, G., Iwasaki, A., and Krishnaswamy, S. Multiscale PHATE exploration of SARS-CoV-2 data reveals multimodal signatures of disease. bioRxiv, DOI: 10.1101/2020.11.15.383661, 2020.

Lafferty, J. and Lebanon, G. Diffusion kernels on statistical manifolds. *Journal of Machine Learning Research*, 6(5): 129–163, 2005.

Le, T., Yamada, M., Fukumizu, K., and Cuturi, M. Tree-sliced variants of Wasserstein distances. In *Advances in Neural Information Processing Systems 32 (NeurIPS 2019)*, 2019.

Leeb, W. *Topics in Metric Approximation*. PhD thesis, Yale University, 2015.

Leeb, W. The mixed lipschitz space and its dual for tree metrics. *Applied and Computational Harmonic Analysis*, 44(3):584–610, 2018.

Leeb, W. and Coifman, R. Hölder-Lipschitz norms and their duals on spaces with semigroups, with applications to Earth mover's distance. *Journal of Fourier Analysis and Applications*, 22:910–953, 2016.

Liberty, E., Woolfe, F., Martinsson, P.-G., Rokhlin, V., and Tygert, M. Randomized algorithms for the low-rank approximation of matrices. *Proceedings of the National Academy of Sciences*, 104(51):20167–20172, December 2007.

Liu, R., Zou, J., and Balsubramani, A. Learning transport cost from subset correspondence. In *the 8th International Conference on Learning Representations (ICLR 2020)*, 2020.

Lucas, C., , Wong, P., Klein, J., Castro, T. B. R., Silva, J., Sundaram, M., Ellingson, M. K., Mao, T., Oh, J. E., Israelow, B., Takahashi, T., Tokuyama, M., Lu, P., Venkataraman, A., Park, A., Mohanty, S., Wang, H., Wyllie, A. L., Vogels, C. B. F., Earnest, R., Lapidus, S., Ott, I. M., Moore, A. J., Muenker, M. C., Fournier, J. B., Campbell, M., Odio, C. D., Casanovas-Massana, A., Herbst, R., Shaw, A. C., Medzhitov, R., Schulz, W. L., Grubaugh, N. D., Cruz, C. D., Farhadian, S., Ko, A. I., Omer, S. B., and Iwasaki, A. Longitudinal analyses reveal immunological misfiring in severe COVID-19. *Nature*, 584:463–469, 2020.

Moon, K. R., Stanley, J. S., Burkhardt, D., van Dijk, D., Wolf, G., and Krishnaswamy, S. Manifold learning-based methods for analyzing single-cell rna-sequencing data. *Current Opinion in Systems Biology*, 7:36–46, 2018.

Moon, K. R., van Dijk, D., Wang, Z., Gigante, S., Burkhardt, D. B., Chen, W. S., Yim, K., van den Elzen, A., Hirn, M. J., Coifman, R. R., Ivanova, N. B., Wolf, G., and Krishnaswamy, S. Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology*, 37(12):1482–1492, 2019.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, É. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(85):2825–2830, 2011.

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

Sato, R., Yamada, M., and Kashima, H. Fast unbalanced optimal transport on a tree. In *Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, pp. 19039–19051, 2020.

Schiebinger, G., Shu, J., Tabaka, M., Cleary, B., Subramanian, V., Solomon, A., Gould, J., Liu, S., Lin, S., Berube, P., Lee, L., Chen, J., Brumbaugh, J., Rigollet, P., Hochedlinger, K., Jaenisch, R., Regev, A., and Lander, E. S. Optimal-transport analysis of single-cell gene expression identifies developmental trajectories in reprogramming. *Cell*, 176(4):928–943.e22, 2019.

Shifrut, E., Carnevale, J., Tobin, V., Roth, T. L., Woo, J. M., Bui, C. T., Li, P. J., Diolaiti, M. E., Ashworth, A., and Marson, A. Genome-wide CRISPR screens in primary human T cells reveal key regulators of immune function. *Cell*, 175(7):1958–1971.e15, 2018.

Shirdhonkar, S. and Jacobs, D. W. Approximate Earth mover's distance in linear time. In *the 2008 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2008.

Shuman, D. I., Vandergheynst, P., and Frossard, P. Chebyshev polynomial approximation for distributed signal processing. In *the 2011 International Conference on Distributed Computing in Sensor Systems and Workshops (DCOSS)*, pp. 1–8, 2011.

Solomon, J., de Goes, F., Peyré, G., Cuturi, M., Butscher, A., Nguyen, A., Du, T., and Guibas, L. Convolutional wasserstein distances: Efficient optimal transportation on geometric domains. *ACM Transactions on Graphics*, 34 (4):1–11, 2015.

Tong, A. and Krishnaswamy, S. Interpolating optimal transport barycenters of patient manifolds, 2020.

Tong, A., Huang, J., Wolf, G., van Dijk, D., and Krishnaswamy, S. TrajectoryNet: A dynamic optimal transport network for modeling cellular dynamics. In *Proceedings of the 37th International Conference on Machine Learning (ICML 2020)*, volume 119 of *PMLR*, pp. 9526–9536, 2020.

Trefethen, L. N. *Approximation Theory and Approximation Practice*. Society for Industrial and Applied Mathematics (SIAM), 2013.

van der Maaten, L. and Hinton, G. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9(86): 2579–2605, 2008.

Yang, K. D. and Uhler, C. Scalable unbalanced optimal transport using generative adversarial networks. In *the 7th International Conference on Learning Representations (ICLR 2019)*, 2019.