Hyperbolic Diffusion Embedding and Distance for Hierarchical Representation Learning

Ya-Wei Eileen Lin [1] **Ronald R. Coifman** [2] **Gal Mishne** [3] **Ronen Talmon** [1]

Abstract

Finding meaningful representations and distances of hierarchical data is important in many fields. This paper presents a new method for hierarchical data embedding and distance. Our method relies on combining diffusion geometry, a central approach to manifold learning, and hyperbolic geometry. Specifically, using diffusion geometry, we build multi-scale densities on the data, aimed to reveal their hierarchical structure, and then embed them into a product of hyperbolic spaces. We show theoretically that our embedding and distance recover the underlying hierarchical structure. In addition, we demonstrate the efficacy of the proposed method and its advantages compared to existing methods on graph embedding benchmarks and hierarchical datasets.

1. Introduction

Hierarchical data is prevalent in many fields of applied science and engineering. Therefore, finding meaningful representations and distances of hierarchical data is an important scientific task. Hyperbolic geometry provides a powerful tool for this purpose; due to their exponential growth, hyperbolic spaces can naturally represent data with tree-like structures (Sarkar, 2011). Indeed, an abundance of methods that involve hyperbolic geometry have been developed in recent years. Notable examples include optimization-based methods (Chamberlain et al., 2017; Nickel & Kiela, 2017; 2018; Chami et al., 2020) and combinatorial methods (Sala et al., 2018; Sonthalia & Gilbert, 2020), to mention just a few. Such hyperbolic geometry-based methods have been successfully applied to central scientific tasks involving hierarchical data in a broad range of fields, e.g., natural language processing (Tifrea et al., 2018), social networks (Verbeek & Suri, 2014), computer vision (Khrulkov et al., 2020), information retrieval (Tay et al., 2018), bioinformatics (Ding & Regev, 2021; Lin et al., 2021), and reinforcement learning (Cetin et al., 2022).

Despite their growing utility, existing methods for hyperbolic representation learning suffer from several notable shortcomings. For example, methods based on optimization of some objective function do not necessarily guarantee optimal representation in terms of standard definitive quality metrics of tree-like geometries (Sonthalia & Gilbert, 2020) and could suffer from numerical instabilities (Mishne et al., 2022). In the context of this work, perhaps the most restrictive disadvantage of many methods for hyperbolic representation learning is that they require the tree graph or tree distance to be known in advance. However, often in practice, we are only given observational data without any prior information, and the underlying hierarchical structures need to be recovered from the ground up.

In this paper, we present a new method for hierarchical data embedding and distance recovery that provably reveals the underlying tree-like structure. In contrast to existing methods, our approach can also be applied to observational data without any prior knowledge of the underlying hierarchical structure. Our method builds on diffusion geometry (Coifman & Lafon, 2006), which is a mathematical framework that facilitates the analysis of high-dimensional data points by capturing their underlying geometric structures. The basic idea behind diffusion geometry is to analyze the similarity between the data points through diffusion propagation. It is based on the construction of a diffusion operator from observations, which is tightly related to the heat kernel and the Laplacian of the underlying manifold. Diffusion geometry has been mainly used for manifold learning, giving rise to multi-scale low-dimensional representations and informative distances. In the past two decades, it has been shown useful in a large number of applications from a broad range of fields, for example, spectral clustering (Nadler et al., 2005), signal processing (Talmon et al., 2013), multi-view dimensionality reduction (Lindenbaum et al., 2020), dynamical systems (Talmon & Coifman, 2013), anomaly detection (Mishne & Cohen, 2012), datasets alignment (Shnitzer et al.,

[1]Viterbi Faculty of Electrical and Computer Engineering, Technion, Haifa, Israel [2]Department of Mathematics, Yale University, New Haven, CT, USA [3]Halicioğlu Data Science Institute, University of California San Diego, La Jolla, CA, USA. Correspondence to: Ya-Wei Eileen Lin <lin.ya-wei@campus.technion.ac.il>.

Proceedings of the 40th International Conference on Machine Learning, Honolulu, Hawaii, USA. PMLR 202, 2023. Copyright 2023 by the author(s).

2022), variational autoencoders (Li et al., 2020), graph analysis (Cheng et al., 2019), hyperspectral image clustering (Murphy & Maggioni, 2019), and non-rigid shape recognition (Bronstein et al., 2010), to name but a few.

We propose an embedding of high-dimensional observations and the corresponding distance between observations that extract their underlying hierarchical structure. Our method conceptually consists of four steps. First, we build a collection of diffusion operators at multiple scales, aiming to reveal the underlying structure of the data. Second, we embed each data point using the diffusion operator at each scale into the Poincaré half-space, which is a particular hyperbolic space model. Third, for each data point, we consider the collection of its embedded points at all scales as a point in the product manifold of hyperbolic spaces, whose mutual relationships are designed to recover the hierarchical structure. We term this representation in the product manifold *hyperbolic diffusion embedding* (HDE). Last, we present the *hyperbolic diffusion distance* (HDD), which naturally stems from the ℓ_1 distance in the product manifold and the Riemannian distance of the hyperbolic space. While existing methods of hyperbolic representation learning, e.g., (Nickel & Kiela, 2017; 2018), embed each high-dimensional data point to a new point in hyperbolic space, our method embeds each data point to a new point in a product of hyperbolic spaces (i.e., a collection of points in hyperbolic spaces), thereby further using the hyperbolic geometry for recovering the hierarchical structure.

We posit that our approach is fundamentally different from existing work. Specifically, our approach combines several components, e.g., diffusion and hyperbolic geometries, for the first time, to the best of our knowledge. At first glance, the combination might seem arbitrary, yet, we show theoretically that HDD is equivalent to the underlying hidden hierarchical distance and that each component has a critical contribution to this result. In addition, we demonstrate the applicability of HDD to hierarchical graph benchmarks, single-cell RNA-sequencing data, and unsupervised hierarchical metric learning tasks. We show that HDD, compared to existing baselines and deep learning-based methods, achieves improved empirical results, both in terms of two standard quality measures of hierarchical representations and in terms of the accuracy of downstream classification. Furthermore, we demonstrate that HDD requires shorter run times than most of the competing methods.

Our main contributions are as follows. First, we present a new method for hierarchical data embedding and distance recovery. Our method can receive only observational data as input without prior knowledge. It is purely data-driven, efficient, and theoretically grounded, and it does not rely on deep learning, optimization, or combinatorial considerations. Second, we propose to combine, for the first time to the best of our knowledge, hyperbolic and diffusion geometries. We exploit this combination and propose to embed data points through their diffusion operators to hyperbolic spaces in order to build a meaningful multi-scale distance metric. Multi-scale distance metrics have been explored in the past using functions defined on Haar wavelet bases (Gavish et al., 2010), partition trees (Mishne et al., 2016; 2017), and dual manifolds (Mishne et al., 2019). Here, we show that using hyperbolic spaces, our multi-scale metrics are capable of taking into account multiple scales of the observational data, from the finest to the coarsest, which plays a key role in the recovery of the hierarchical structure. Third, we showcase improved performance compared to leading recent baselines on several benchmarks, demonstrating accurate and efficient hierarchical structure extraction.

2. Background

Diffusion Geometry. Diffusion geometry (Coifman & Lafon, 2006) is a framework for high-dimensional data analysis. It is based on revealing similarities between the data points by constructing multi-scale "diffusion" processes. Under the manifold assumption (Fefferman et al., 2016), i.e., assuming that the high-dimensional data lie on a low-dimensional manifold, diffusion geometry facilitates the recovery of the underlying manifold. Below, we outline the main steps of the construction of diffusion geometry that we utilize in our methods.

Let $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$ be a set of data points in an ambient space $\mathbb{R}^m$ that lie on a hidden manifold. Let $\mathbf{W}$ be a pairwise affinity matrix, given by

$$\mathbf{W}(i,i') = \exp\left(-d^2(i,i')/\epsilon\right), \tag{1}$$

where $d(\cdot,\cdot)$ represents a suitable distance between the data points $\mathbf{x}_i$ and $\mathbf{x}_{i'}$, and ϵ is a tunable kernel scale parameter, which in practice is often set as the median of distances multiplied by a constant or adjusted according to the nearest neighbors (Zelnik-Manor & Perona, 2004; Keller et al., 2009; Ding & Wu, 2020). The set $\mathcal{X}$ and the affinity matrix $\mathbf{W}$ form an undirected weighted graph, where $\mathcal{X}$ is the node set and $\mathbf{W}$ is the edges' weight matrix. By normalizing the affinity matrix twice as follows

$$\widetilde{\mathbf{W}} = \mathbf{S}^{-1}\mathbf{W}\mathbf{S}^{-1}, \qquad \mathbf{S}(i,i) = \textstyle\sum_j \mathbf{W}(i,j), \tag{2}$$

$$\mathbf{P} = \mathbf{D}^{-1}\widetilde{\mathbf{W}}, \qquad \mathbf{D}(i,i) = \textstyle\sum_j \widetilde{\mathbf{W}}(i,j), \tag{3}$$

the resulting matrix $\mathbf{P}$ is viewed as a transition probability matrix of a Markov chain defined on the graph.

The matrix $\mathbf{P}$ is termed *diffusion operator* since it can be used to propagate mass between nodes on the graph. Let $\mathbf{p}_i^t = \mathbf{P}^t e_i \in \mathbb{R}^n$ be the density on the graph after t diffusion propagation steps, where $e_i \in \mathbb{R}^n$ is the indicator vector of the i-th node, and $t \in (0,1]$ is the diffusion time.

Note that by definition, $\mathbf{p}_i^t$ is a well-defined discrete distribution on the graph because $\mathbf{p}_i^t(j) \geq 0$ for any $j = 1, \ldots, n$, and $\sum_{j=1}^n \mathbf{p}_i^t(j) = 1$. Note that considering multiple diffusion times t gives rise to a collection of multi-scale densities for each data point. This construction provides a family of multi-scale distances and embeddings, called diffusion distance and diffusion maps, respectively (see Appendix A for more details).

The diffusion operator recovers the underlying manifold in the following sense (Coifman & Lafon, 2006). In the limit $n \to \infty$ and $\epsilon \to 0$, the operator $\mathbf{P}^{t/\epsilon}$ converges to the Neumann heat kernel of the underlying manifold, given by $\mathbf{H}_t = \exp(-t\Delta)$, where Δ is the Laplace–Beltrami operator on the manifold. In other words, the diffusion operator (matrix) is a discrete approximation of the heat kernel on the manifold.

Poincaré Half-Space Model of Hyperbolic Space. Hyperbolic geometry is a non-Euclidean geometry with constant negative curvature. In this work, we consider the n-dimensional Poincaré half-space model of hyperbolic space with curvature -1 (Beardon, 2012). It is defined by $\mathbb{H}^n = \{\mathbf{x} \in \mathbb{R}^n | \mathbf{x}(n) > 0\}$ with the Riemannian metric tensor $ds^2 = \frac{d\mathbf{x}^2(1)+d\mathbf{x}^2(2)+\ldots+d\mathbf{x}^2(n)}{\mathbf{x}^2(n)}$. Given two points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{H}^n$, the Riemannian distance is computed by

$$d_{\mathbb{H}^n}(\mathbf{x}, \mathbf{y}) = 2\sinh^{-1}\left(\frac{\|\mathbf{x}-\mathbf{y}\|_2}{2\sqrt{\mathbf{x}(n)\mathbf{y}(n)}}\right), \tag{4}$$

where $\|\cdot\|_2$ is the Euclidean norm.

Product Manifolds and Distances. Product manifolds (Turaga & Srivastava, 2016) provide a product space for mixed curvature representation learning (Gu et al., 2018; Skopek et al., 2020). Consider a set of Riemannian manifolds denoted by $\{(\mathcal{M}_l, g_l)\}_{l=1}^L$. The product manifold is defined by the Cartesian product $\mathcal{M} = \mathcal{M}_1 \times \mathcal{M}_2 \times \ldots \times \mathcal{M}_L$, whose dimension is the sum of the dimensions of the factor manifolds $\mathcal{M}_l$. The product manifold $\mathcal{M}$ is equipped with the Riemannian metric tensor $g = \sum_l g_l$ (Ficken, 1939). Different distances can be considered in $\mathcal{M}$. In this work, we use the ℓ_1 distance and show that this choice enables the recovery of the underlying hierarchical structure. The ℓ_1 is defined by

$$d_{\mathcal{M}}^{\ell_1}(x, y) = \sum_{l=1}^{L} d_{\mathcal{M}_i}(x^l, y^l), \tag{5}$$

where $x = (x^1, \ldots, x^L), y = (y^1, \ldots, y^L) \in \mathcal{M}$ such that $x^l, y^l \in \mathcal{M}_i$, and $d_{\mathcal{M}_l}$ is the geodesic distance on $\mathcal{M}_l$.

3. Problem Formulation

The problem of learning hierarchical representations can be formulated in both graph embedding context, e.g., as in (Nickel & Kiela, 2017; Sala et al., 2018; Sonthalia & Gilbert, 2020), and in hierarchical distance recovery context, e.g., as in (Dasgupta, 2016; Klimovskaia et al., 2020; Chami et al., 2020; Sonthalia & Gilbert, 2020; Fang et al., 2021). Their settings and goals are presented below.

In the *graph embedding* context, a tree graph $G = (\mathcal{X}, \mathcal{E}, \mathbf{W})$ is given, where $\mathcal{X}$ is the vertex set with n nodes, $\mathcal{E}$ is the edge set with the weight matrix $\mathbf{W}$. In this case, the vertex set $\mathcal{X}$ can be viewed as a discrete subset of a hierarchical metric space $(\mathcal{T}, d_{\mathcal{T}})$, where the hierarchical distance $d_{\mathcal{T}}$ on the tree nodes $\mathcal{X}$ coincides with the shortest path distance on the graph. Here, the objective is to find a node embedding into a metric space whose distance approximates the distance $d_{\mathcal{T}}$.

In practice, we are often given solely observational data in a high-dimensional ambient space that is assumed to have a latent hierarchical structure that is not explicitly given. Therefore, our primary focus in this work is on the *hierarchical distance recovery* context, where, given a set of n data points $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$, we view these as a node set of a hidden (tree) graph. Typically, the points are embedded in some high-dimensional ambient Euclidean space, i.e., $\mathbf{x}_i \in \mathbb{R}^m$, and they have an underlying hierarchical structure. This is often formulated by assuming that the points lie in a hidden hierarchical metric space $(\mathcal{T}, d_{\mathcal{T}})$, where both the space $\mathcal{T}$ and the hierarchical distance $d_{\mathcal{T}}$ are inaccessible. Given the observations $\mathcal{X}$, the goal in hierarchical distance recovery is to find a distance that approximates the hidden hierarchical distance $d_{\mathcal{T}}$.

4. Hyperbolic Diffusion Distance

We begin by considering our primary problem setting of hierarchical distance recovery. Let $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$ be the data set described above. We follow the construction of the diffusion geometry described in Section 2. First, we build the matrix $\mathbf{W}$ based on a Gaussian kernel as in Eq. (1), where d is a suitable distance in the ambient space $\mathbb{R}^m$. Then, the diffusion operator $\mathbf{P}$ is constructed according to Eqs. (2) and (3). Note that this procedure implicitly constructs a graph $G = (\mathcal{X}, \mathcal{E}, \mathbf{W})$, where $\mathcal{X}$ is viewed as the vertex set and $\mathbf{W}$ is the weight matrix of the edges. Consequently, $\mathbf{P}$ is a transition probability matrix of a random walk on this graph (Coifman & Lafon, 2006). We present an illustration of the high-dimensional data set $\mathcal{X}$ and the underlying hierarchical structure in Fig. 1(a).

This graph viewpoint associates the hierarchical distance recovery and graph embedding contexts. Namely, in the case of graph embedding, the tree graph $G = (\mathcal{X}, \mathcal{E}, \mathbf{W})$ is given,

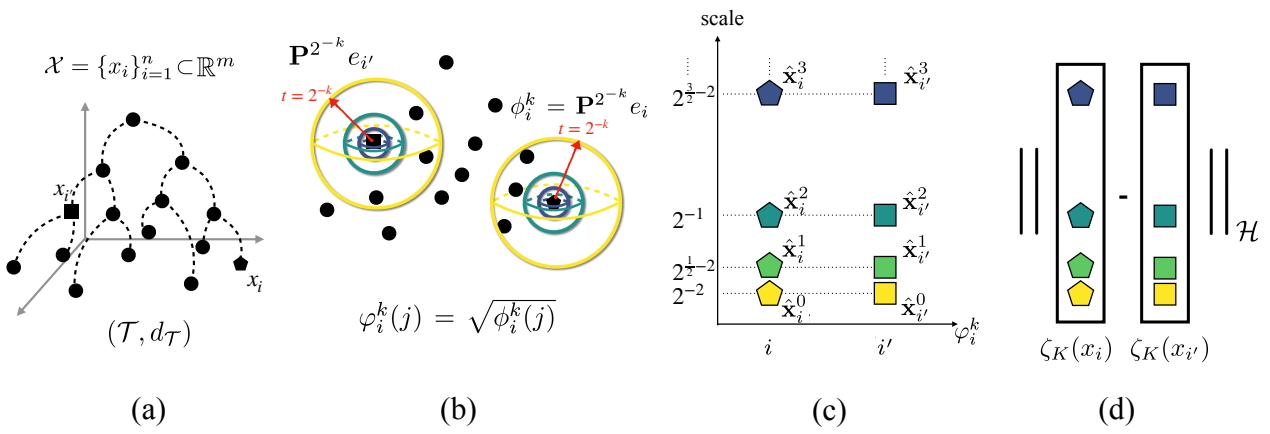

Figure 1. An illustration of HDE and HDD. (a) Given a dataset $\mathcal{X} = \{x_i\}_{i=1}^n \subset \mathbb{R}^m$ with an underlying tree-like structure. (b) Build a diffusion operator $\mathbf{P}$ by connecting neighboring points. Consider multiple scales of the operator on a dyadic grid $\{\mathbf{P}^{2^{-k}}\}_{k=0}^K$ for $K \in \mathbb{Z}_0^+$. For each point x_i, define a collection of propagated densities $\phi_i^k = \mathbf{P}^{2^{-k}} e_i$. Here, two densities that are propagated from two points x_i and $x_{i'}$ are shown at various diffusion times on a dyadic grid. The different scales are represented by using different colors. (c) HDE is given by $\zeta_K : \mathcal{X} \to \mathcal{H}$, which maps a point x_i into a product of hyperbolic spaces $\mathcal{H} = \mathbb{H}^{n+1} \times \ldots \times \mathbb{H}^{n+1}$ using $\{\varphi_i^k(j) = \sqrt{\phi_i^k(j)}\}_{k=0}^K$. The resulting multi-scale HDE of the two points x_i and $x_{i'}$ are presented, where the colors correspond to the colors in (b). Here, the square root of the propagated densities is shown on the x-axis, and the scales $2^{k\alpha-2}$ of the HDE are shown on the y-axis. (d) HDD is defined by the ℓ_1 distance on the product of hyperbolic spaces $\mathcal{H}$.

and the diffusion operator $\mathbf{P}$ is defined as an approximation of the heat kernel $\mathbf{H} = \exp(-\mathbf{L})$, as described in Section 2. Therefore, from this point on, the two contexts coincide, and the proposed method is applicable to both settings.

We consider a dyadic grid of diffusion times $t = 2^{-k}$ for $k \in \mathbb{Z}_0^+$. Fig. 1(b) displays the propagated densities at several diffusion times $t = 2^{-k}$. For convenience, let $\phi_i^k, \varphi_i^k \in \mathbb{R}^n$ denote $\phi_i^k = \mathbf{P}^{2^{-k}} e_i$ and $\varphi_i^k(j) = \sqrt{\phi_i^k(j)}$ for $j = 1, \ldots, n$.

First, we propose to embed the point $\mathbf{x}_i$ using the diffusion operator $\mathbf{P}^{2^{-k}}$, denoted by the pair (i, k), into the Poincaré half-space model $\mathbb{H}^{n+1}$ by concatenating φ_i^k with a function of the diffusion time 2^{-k} as follows:

$$(i, k) \mapsto \hat{\mathbf{x}}_i^k = [(\varphi_i^k)^\top, 2^{k\alpha-2}]^\top \in \mathbb{H}^{n+1}, \qquad (6)$$

where $0 < \alpha < 1$ is a parameter that scales the diffusion propagation. We remark that the factor 2^{-2} added to the diffusion time in Eq. (6) is an important weight term in the heat kernel approximation, allowing to appropriately capture the local intrinsic association between the hierarchies at each diffusion time scale $t = 2^{-k}$.

Next, we extend the embedding by considering multiple diffusion times $\{2^{-k}\}_{k=0}^K$ simultaneously, where $K \in \mathbb{Z}_0^+$ denotes the maximal scale. The embedding space is the product manifold $\mathcal{H} = \mathbb{H}^{n+1} \times \mathbb{H}^{n+1} \times \ldots \times \mathbb{H}^{n+1}$ of $(K+1)$ elements (i.e., $\mathcal{H} \subset \mathbb{R}^{(n+1)(K+1)}$), and the multi-scale HDE is a function $\zeta_K : \mathcal{X} \to \mathcal{H}$ defined by

$$\zeta_K(x_i) = \left[(\hat{\mathbf{x}}_i^0)^\top, (\hat{\mathbf{x}}_i^1)^\top, (\hat{\mathbf{x}}_i^2)^\top, \ldots, (\hat{\mathbf{x}}_i^K)^\top\right]^\top. \qquad (7)$$

Fig. 1(c) illustrates the multi-scale HDE of two points, $\zeta_K(x_i)$ and $\zeta_K(x_{i'})$, where $K = 4$. We used the Poincaré half-space model because of its natural representation of the diffusion time on a dyadic grid (i.e., $t = 2^{-k}$ for $k \in \mathbb{Z}_0^+$) as well as its capability to represent multiple diffusion time scales simultaneously.

Note that considering such a multi-scale embedding significantly departs from the common practice. Existing methods of hierarchical representation learning in the Poincaré model, e.g., (Nickel & Kiela, 2017; Chami et al., 2020), learns the hyperbolic embedding by an optimization method that pushes points toward the boundary of the Poincaré model or views the data points as tree leaves and attempts to place their embedding close to the boundary of the Poincaré model. In contrast, our method embeds each data point as a collection of points in hyperbolic spaces, thereby further exploiting the negative curvature of the space. Specifically, by construction, when the diffusion time goes to zero, the diffused densities concentrate at single points, i.e., $\mathbf{P}^t e_i \xrightarrow{t \to 0} e_i$. This implies that when the scale k is large (blue in Fig. 1), the propagated densities ϕ_i^k provide a local view of the data, and by Eq. (6), the embedded points at scale k are pushed toward the upper part of the Poincaŕe half space, giving rise to a fine-scale distance. Conversely, when k is small (yellow in Fig. 1), we have a coarse-grained view

Algorithm 1 Hyperbolic Diffusion Embedding and Distance

Input: Diffusion operator $\mathbf{P} \in \mathbb{R}^{n\times n}$, parameter α, and maximum scale K
Output: Hyperbolic diffusion distance $d_{\text{HDD}}(i,i')$ for all $i,i' \in [1,n]$ and embedding $\hat{\mathbf{X}} \in \mathbb{R}^{n\times((n+1)(K+1))}$

$\mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top = \texttt{eig}(\mathbf{P})$
$k \leftarrow 0$
while $k \leq K$ **do**
  $\hat{\mathbf{\Lambda}}_k \leftarrow \mathbf{\Lambda}^{2^{-k}}$
  $\hat{\mathbf{X}}_k \leftarrow [\sqrt{(\mathbf{U}\hat{\mathbf{\Lambda}}_k\mathbf{V}^\top)^\top}, 2^{k\alpha-2}\mathbf{1}_n]^\top$
  $k \leftarrow k+1$
end while
$\hat{\mathbf{X}} \leftarrow [\hat{\mathbf{X}}_0^\top, \hat{\mathbf{X}}_1^\top, \ldots, \hat{\mathbf{X}}_K^\top]^\top$
for $i,i' \in \{1,2,\ldots,n\}$ **do**
  $d_{\text{HDD}}(i,i') \leftarrow \sum_{k=0}^{K} 2\sinh^{-1}\left(2^{-k\alpha+1}\left\|\hat{\mathbf{x}}_i^k - \hat{\mathbf{x}}_{i'}^k\right\|_2\right)$
end for

of the data, and in this case, the proposed embedding gives rise to exponentially scaled distances, which are consistent with the scaling of a tree distance. This argument is made formal in the following statement.

Proposition 1. *There is a constant $0 < C < 1$ such that for any $x_i, x_{i'} \in \mathcal{X}$ and $k_1 \leq k_2$, $k_1, k_2 \in \mathbb{Z}_0^+$, we have*

$$C\cdot 2^{-(k_2-k_1)\alpha} \leq \frac{d_{\mathbb{H}^{n+1}}(\hat{\mathbf{x}}_i^{k_2}, \hat{\mathbf{x}}_{i'}^{k_2})}{d_{\mathbb{H}^{n+1}}(\hat{\mathbf{x}}_i^{k_1}, \hat{\mathbf{x}}_{i'}^{k_1})} \leq \frac{1}{C}\cdot 2^{-(k_2-k_1)\alpha}. \quad (8)$$

The proof of Proposition 1 is in Appendix B.

In terms of dimensionality, each diffusion operator embeds a point into an n-dimensional density, which in turn is mapped to $\mathbb{H}^{n+1}$. Then, considering the product space of $(K+1)$ scales overall results in an $(n+1)(K+1)$-dimensional HDE. This potential dimensionality increase also departs from common practice that typically aims at dimension reduction. However, as we show in Section 5, the gain is the ability to recover the underlying hierarchical structure.

Finally, we propose a new distance called the Hyperbolic Diffusion Distance (HDD), using the ℓ_1 distance on the product manifold $\mathcal{H}$, which is defined in Eqs. (4) and (5). Formally, the HDD between two points $x_i, x_{i'} \in \mathcal{X}$ is defined by

$$\begin{aligned} d_{\text{HDD}}(i,i') &:= d_{\mathcal{H}}^{\ell_1}\left(\zeta_K(x_i), \zeta_K(x_{i'})\right) \\ &= \sum_{k=0}^{K} 2\sinh^{-1}\left(2^{-k\alpha+1}\left\|\varphi_i^k - \varphi_{i'}^k\right\|_2\right), \end{aligned} \quad (9)$$

where α is the same parameter as in Eq. (6). Note that the function $\sinh^{-1}$ in Eq. (9) arises from the Riemannian distance in the hyperbolic space $\mathbb{H}^{n+1}$ at each scale. In practice, it attenuates large distance values. Fig. 1(d) illustrates this multi-scale HDD. We summarize the construction of the proposed HDE and HDD in Algorithm 1.

5. Theoretical Justification

Here we show that the proposed HDE and HDD are theoretically grounded. The intuition underlying the constructions of HDE and HDD in Algorithm 1 is as follows. The diffusion operator $\mathbf{P}^t$ is designed to reveal the local connectivity at diffusion time scale t. Considering multiple scales in a dyadic grid associates different diffusion timescales, namely, neighborhoods of different sizes. The multi-scale embedding of the corresponding diffusion operators in hyperbolic space naturally endows a hierarchical relationship between the diffusion timescales, enabling the recovery of the underlying hierarchical structure. Next, we make this intuitive explanation formal.

Theorem 1. *For $0 < \alpha < \frac{1}{2}$ and sufficiently large K and n, the hyperbolic diffusion distance d_{HDD} is equivalent to $d_{\mathcal{T}}^{2\alpha}$.*

Theorem 1 implies that the proposed HDD recovers the underlying hierarchical distance even when $d_{\mathcal{T}}$ is not given or when we do not have access to the explicit tree structure. In addition, Theorem 1 suggests that in practice, the parameter α should be set to be close to $\frac{1}{2}$ so that HDD approximates the hierarchical distance. That is, for $\alpha \to \frac{1}{2}$, the obtained d_{HDD} is approximately 0-hyperbolic (Gromov, 1987).

Theorem 1 is stated under the assumption that $(\mathbf{P}^t)_{t\in(0,1]}$ is a point-wise approximation of the heat kernel, namely, $\mathbf{P}^t(i,i') \approx a_t(x_i, x_{i'})$, where $a_t(\cdot,\cdot)$ is a heat kernel (Grigoryan, 2009). Such an approximation, mainly in the limit $n \to \infty$ and $\epsilon \to 0$, was shown and studied in (Coifman & Lafon, 2006; Singer, 2006; Belkin & Niyogi, 2008). In addition, three strong regularity conditions are required for the above metric recovery result. Importantly, the heat kernel admits these conditions (see Appendix C).

The first condition is an *upper bound* on the operator elements. There is a non-negative and monotonic decreasing function $f_1 : \mathbb{R}_+ \to \mathbb{R}$ and a number $\beta > 0$ such that for any $\gamma < \beta$, we have $\int_{\mathbb{R}_+} \tau^{3n+\gamma} f_1(\tau) d\tau/\tau < \infty$. The square root of the operator elements for all $t \in (0,1]$ is then upper-bounded by

$$\sqrt{\mathbf{P}^t(i,i')} \leq \frac{1}{t^{\frac{3n}{2\beta}}} f_1\left(\frac{d_{\mathcal{T}}(x_i, x_{i'})}{t^{\frac{1}{\beta}}}\right). \quad (10)$$

The second condition is a *lower bound* on the operator elements. There is a monotonic decreasing function $g_1 : \mathbb{R}_+ \to \mathbb{R}$ and $R > 0$ such that for all $t \in (0,1]$ and all $d_{\mathcal{T}}(x_i, x_{i'}) < R$, the square root of the operator elements

is lower-bounded by

$$\sqrt{\mathbf{P}^t(i,i')} \geq \frac{1}{t^{\frac{n}{2\beta}}} g_1 \left(\frac{d_{\mathcal{T}}(x_i, x_{i'})}{t^{\frac{1}{\beta}}} \right). \tag{11}$$

The third condition is a *Hölder continuity condition*. There is a constant $\Theta > 0$ sufficiently small such that for all $t \in (0,1]$, all $x_i, x_{i'} \in \mathcal{X}$ with $d_{\mathcal{T}}(x_i, x_{i'}) \leq t^{\frac{1}{\beta}}$ and all $x_j \in \mathcal{X}$, the element value of the Hellinger measure (Hellinger, 1909) is upper-bounded by

$$|\sqrt{\mathbf{P}^t(i,j)} - \sqrt{\mathbf{P}^t(i',j)}|^2 \leq \left(\frac{d_{\mathcal{T}}(x_i, x_{i'})}{t^{\frac{1}{\beta}}} \right)^{2\Theta} \frac{1}{t^{\frac{n}{\beta}}} f_1 \left(\frac{d_{\mathcal{T}}(x_i, x_{i'})}{t^{\frac{1}{\beta}}} \right). \tag{12}$$

Here, we presented only the main results. The proof of Theorem 1 and more details appear in Appendix C.

We conclude this subsection with a couple of remarks. First, our results and derivations rely on and extend the work of (Leeb & Coifman, 2016), who proposed a multi-scale distance based on diffusion geometry and showed that it approximates the geodesic distance on a closed Riemannian manifold with non-negative curvature (Goldberg & Kim, 2012; Leeb, 2015), under the conditions of geometric and semi-groups regularities. However, their approximation does not apply to hierarchical structures, such as trees or tree-like structures, that are negatively curved manifolds, as we empirically demonstrate in Appendix E. Here, as a remedy, motivated by the geometric insights presented in (Leeb & Coifman, 2016), we follow the work of (McKean, 1970; Grigor'yan & Noguchi, 1998; Frank & Kovarik, 2013; Zelditch, 2017) for heat kernels on negative curvature spaces. Second, unlike Euclidean spaces where the product of Euclidean spaces is Euclidean, i.e., $(\mathbb{R}^{k_1})^{k_2} = \mathbb{R}^{k_1 \cdot k_2}$, considering the product of (curved) hyperbolic spaces gives $(\mathbb{H}^{k_1})^{k_2} \neq \mathbb{H}^{k_1 \cdot k_2}$, for $k_1, k_2 \in \mathbb{Z}^+$. Therefore, embedding into the space $\mathbb{H}^{n(K+1)}$ does not generate the same metric as HDD, which admits a canonical Riemannian metric in $\mathcal{H}$. The specific construction of HDD is unique and essential to recover the hierarchy.

6. Experimental Results

We investigate the proposed HDE and HDD in hierarchical graph embedding and distance recovery contexts. Specifically, we apply it to (i) several graphs serving as benchmarks for hierarchical graph embedding, (ii) single-cell gene expression data for the recovery of the hidden hierarchical structure, and (iii) unsupervised hierarchical metric learning tasks. We refer to Appendix D for more details of these experiments and to Appendix E for additional experiments including a toy example and ablation study. The code is available at the link `https://github.com/Ya-Wei0/HyperbolicDiffusionDistance`.

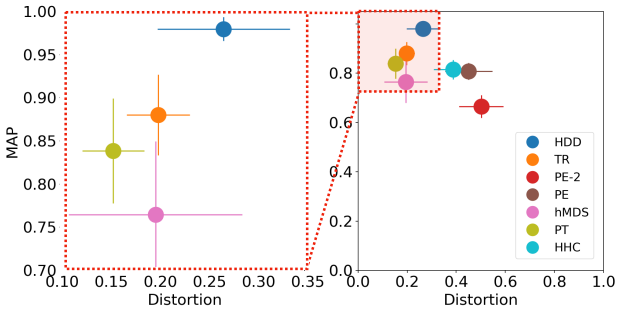

Figure 2. Distortion-MAP results of hierarchical graph embedding. Each point represents the mean over the five benchmark graphs, and the whiskers represent the standard deviation. The larger the MAP and the smaller the average distortion, the better the embedding quality.

6.1. Hierarchical Graph Embedding

We demonstrate our method in the context of hierarchical graph embedding on five benchmark graphs, which were considered in (Sala et al., 2018). These benchmarks include (i) a small and fully-balanced tree consisting of 40 nodes, (ii) a phylogenetic tree consisting of 344 nodes (Sanderson et al., 1994; Hofbauer et al., 2016), (iii) a graph of disease relations consisting of 516 nodes, (iv) a CS-PhD graph of the relations between advisors and PhD students consisting of 1025 nodes (De Nooy et al., 2018), and (v) a general relativity and quantum cosmology arXiv collaboration network with 4185 nodes (Leskovec et al., 2007). These five graphs contain trees, tree-like graphs, and dense graph.

Given each of these graphs, consisting of nodes and edges, our goal here is to find a node embedding $\xi(\cdot)$ in a metric space, where the metric d_ξ represents the hierarchical distance $d_{\mathcal{T}}$ defined by the shortest path on the given graph. For this purpose, we apply Algorithm 1 with $\alpha = \frac{1}{2}$ and $K \in \{0, 1, \ldots, 19\}$. Implementation details are described in Appendix D.

We compare our method to five hierarchical embedding methods. The first is the tree representation (TR) obtained by a divide-and-conquer tree construction (Sonthalia & Gilbert, 2020). The second is the Poincaré embedding (PE) (Nickel & Kiela, 2017), which is a neural network for graph embedding that learns the hyperbolic representation in the Poincaré model. Two additional methods are taken from (Sala et al., 2018): a PyTorch (PT) implementation of an SGD-based algorithm optimized over a principal geodesic analysis loss function, and hyperbolic multi-dimensional scaling (hMDS), which takes a pairwise distance matrix as input and returns an embedding in hyperbolic space that best represents the input distances. The fifth is the hyperbolic embedding obtained by the hyperbolic hierarchical clustering (HHC) (Chami et al., 2020) using a continuous relaxation of Dasgupta's cost (Dasgupta, 2016). Following

the experiment protocol in (Sala et al., 2018), we test embedding spaces of dimensions 2, 5, 10, 50, 100, and 200 for the PE, hMDS, and PT methods, and the best results are reported. In addition, following the common practice, we also present the results of the PE into the 2-dimensional Poincaré disk and denote it by PE-2.

To quantitatively evaluate the obtained embeddings, we use two commonly-used fidelity measures. The first is the mean average precision (MAP), which is given by

$$\frac{1}{|\mathcal{X}|}\sum_{x_i\in\mathcal{X}}\frac{1}{d(x_i)}\sum_{x_{i_j}\in N_{x_i}}\frac{|N_{x_i}\cap B(\xi(x_i),\xi(x_{i_j}))|}{B(\xi(x_i),\xi(x_{i_j}))}, \quad (13)$$

where $N_{x_i}=\{x_{i_1},\ldots,x_{i_{d(x_i)}}\}$ is the set of the neighborhood of x_i in the given graph, $d(x_i)$ is the degree of x_i in the given graph, and $B(x,x_{i_j})=\{y|d_\xi(\xi(x_i),\xi(y))\leq d_\xi(\xi(x_i),\xi(x_{i_j}))\}$ is the set of points within the smallest ball that is centered at $\xi(x)$ and contains x_{i_j} in the embedded space. The second measure is the average distortion given by

$$\frac{1}{\binom{|\mathcal{X}|}{2}}\sum_{x_i\neq x_{i'}\in\mathcal{X}}\frac{|d_\xi(\xi(x_i),\xi(x_{i'}))-d_\mathcal{T}(x_i,x_{i'})|}{d_\mathcal{T}(x_i,x_{i'})}. \quad (14)$$

Note that the closer the MAP is to 1, the better the embedding distance d_ξ locally preserves the desired hierarchical distance $d_\mathcal{T}$. In addition, the smaller the average distortion is, the larger the (global) similarity between d_ξ and $d_\mathcal{T}$ is.

Fig. 2 presents the MAP and distortion obtained by Algorithm 1 and the competing methods. For brevity, we present here the mean and standard deviation over the five benchmarks, and the results for each benchmark separately appear in Appendix E. We see that compared to the baselines, HDD presents a trade-off. It yields the best MAP with a small standard deviation, yet, its obtained distortion is larger than TR, hMDS, and PT. Note that HDD is strictly better in terms of MAP and distortion than the popular PE in two or more dimensions. We report the run time and stability in Appendix E, showing that HDD takes a remarkably shorter computational time than PE, PT, and HHC. While HDD requires a longer run time than TR, the advantage of HDD over TR in terms of MAP is significant, as depicted in Fig. 2.

6.2. Single-Cell Gene Expression Data

We examine HDD in the context of hierarchical distance recovery. In contrast to the graph embedding task, this context fully exploits the use of the proposed multi-scale diffusion geometry that is designed to reveal the hidden hierarchical (tree) structure. For this purpose, we consider single-cell RNA sequencing (scRNA-seq) data (Tanay & Regev, 2017). It is argued that single-cell development can be well modeled using hierarchical representations (duVerle et al., 2016), providing biological insights into cell developmental trajectories and disease progression (van Galen et al., 2019). Revealing the latent hierarchies underlying the cell types is a key task in differentiating genetic treatments and immune responses, which are useful for further biological tasks.

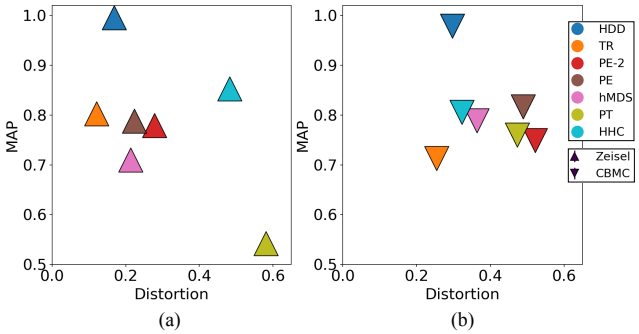

Figure 3. Distortion-MAP results of scRNA-seq datasets: (a) Zeisel and (b) CBMC. The larger the MAP and the smaller the average distortion, the better the hierarchy distance recovery.

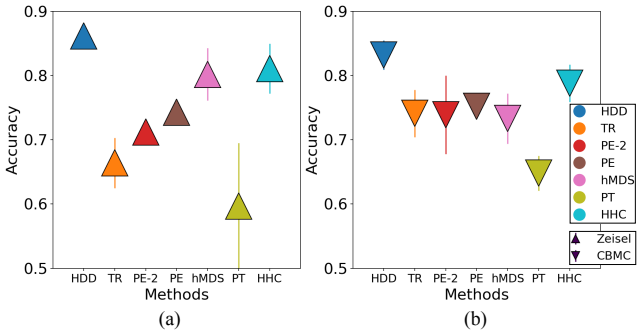

Figure 4. Classification accuracy of scRNA-seq datasets: (a) Zeisel and (b) CBMC.

We test two scRNA-seq datasets taken from (Dumitrascu et al., 2021): (i) the mouse cortex and hippocampus dataset (Zeisel) consisting of 3005 single-cells with seven cell types and 4000 gene markers (Zeisel et al., 2015), and (ii) the cord blood mononuclear cell study (CBMC) comprising 8617 single-cells with 13 cell types and 500 gene markers (Stoeckius et al., 2017).

The single cells are viewed as samples in a high-dimensional ambient space, where the genes are viewed as features. Given these data, we apply Algorithm 1, obtaining an embedding of the single cells into a hierarchical metric space. We compare our method to the same baselines as in Section 6.1. A distance based on the cosine similarity computed in the ambient space of the scRNA-seq data expression levels is used as the input distance for Algorithm 1 and the competing methods. We note that the choice of distance in the original space is critical. Here, we employed a distance

Table 1. Classification accuracy of Zoo, Iris, Glass, and Segmentation datasets (highest accuracy in bold and second highest underlined).

Dataset (#Points, #Classes)		HDD	TR	PE-2	PE	hMDS	PT	HHC
Zoo	(101, 7)	**0.898±0.012**	0.854±0.049	0.779±0.038	0.824±0.018	0.822±0.014	0.842±0.015	0.861±0.044
Iris	(150, 3)	0.883±0.007	0.859±0.021	0.782±0.030	0.846±0.024	0.851±0.010	**0.895±0.009**	0.852±0.019
Glass	(214, 6)	**0.654±0.011**	0.607±0.013	0.503±0.036	0.553±0.027	0.609±0.013	0.556±0.046	0.610±0.007
ImaSeg	(2310, 7)	**0.701±0.024**	0.654 ±0.022	0.599±0.017	0.679±0.038	0.658±0.016	0.641±0.027	0.661±0.017

based on the standard and commonly-used cosine similarity calculated in the ambient space based on prior research (Jaskowiak et al., 2014).

To evaluate the embedding and distance, we use the same quantitative measures as in Section 6.1. To compute these measures, we exploit the fact that a tree graph of the cell types is provided with each dataset. Importantly, this information is used only for evaluation but kept hidden from the distance recovery methods.

Fig. 3 presents the obtained MAP and average distortion of Zeisel and CBMC. We see that HDD obtains the highest MAP values by a large margin and the second-best result in terms of the average distortion, which is very close to the best result. The evaluation of the methods' run times (see Appendix E) shows that HDD is more efficient than PE, PT, and HHC. This suggests that HDD is applicable to large scRNA-seq datasets as well.

To further evaluate the results, we make use of the availability of the cell labels in this dataset to also examine HDD through classification. Specifically, we apply the nearest centroid classifier with the recovered distance metric of each method as input. The reported classification accuracy is obtained by averaging over ten different runs; in each run, the dataset is randomly split into 80% training set and 20% testing set. Fig. 4 displays the mean and the standard deviation of the classification accuracy of Zeisel and CBMC. We see that HDD outperforms the other methods by a large margin for both Zeisel and CBMC datasets. This suggests that HDD extracts well the tree-like structure underlying biological data. Conversely, we see that TR and hMDS, which obtain small distortion in the graph embedding context, do not perform well in this downstream task.

6.3. Unsupervised Hierarchical Metric Learning

We further test the proposed HDD in the context of unsupervised hierarchical metric learning. We consider four datasets from the UCI Machine Learning Repository (Dua & Graff, 2017): (i) the Zoo dataset consisting of 101 data points of seven types of animals with 17 features, (ii) the Iris dataset comprising 150 samples from three kinds of Iris plants with four features, (iii) the Glass dataset containing 214 instances of six classes with 10 features, and (iv) the image segmentation (ImaSeg) dataset consisting of 2310 instances from seven outdoor images with 19 features. These datasets were used in (Chami et al., 2020) for evaluating embedding and clustering in hyperbolic space under the working assumption that they have some degree of underlying hierarchical structures. Here, the evaluation of HDE and HDD is done through downstream classification based on the (dis)similarity of the learned embedding and distance (Jordan & Mitchell, 2015). Such evaluation is affected by the hyperbolicity of the datasets, i.e., it depends on the extent the data adhere to hyperbolic geometry. Since there is no such ground truth hierarchical information (in contrast to the datasets considered in Section 6.2), the δ-hyperbolicity cannot be naively computed. Still, following (Chami et al., 2020), we posit that comparing different methods for hyperbolic embedding and distance on these datasets gives useful information on the ability of HDD to reveal hierarchical structures given only observational data, compared to the competing methods.

The downstream classification is carried out in the same way as in Section 6.2. Algorithm 1 and the baseline methods are applied to these datasets with a distance based on cosine similarity as an input without using any label information. Then, a dissimilarity classification based on the learned hierarchical distance using the nearest centroid classifier is applied. We use cross-validation with ten repetitions, in which the dataset is randomly divided into 80% training set and 20% testing set.

The classification accuracy is presented in Table 1, showing the mean and the standard deviation of the results averaged over ten trials. We see that HDD outperforms the competing methods in three out of the four datasets and obtains the second-best classification accuracy in the remaining dataset. These results further demonstrate that HDD gives rise to a useful distance.

7. Conclusion

We presented a new method for hierarchical data embedding and corresponding distance, termed HDE and HDD, respectively, which can receive as input either a graph or observational data with a hidden hierarchical structure. Our method is primarily based on diffusion geometry that en-

ables us to construct multi-scale propagated densities, which are, in turn, embedded in a product of hyperbolic spaces. We theoretically show that the ℓ_1 distance between the embedded points in this space is equivalent to the tree-like distance in the hierarchical space of the input graph or data. In contrast to the common practice in hierarchical data representation using hyperbolic geometry, our method represents each point as a collection of embedded propagated densities rather than a single point in hyperbolic space. We test HDE and HDD on benchmark graph embedding tasks and on single-cell gene expression data sets, demonstrating significant advantages compared to existing methods in terms of standard quantitative metrics and run time. In addition, we demonstrate that HDD can lead to improved downstream classification accuracy on several benchmarks. Because our method is computationally efficient, not based on optimization or deep learning, and differentiable, we posit that it can potentially be incorporated into various loss functions of a broad range of downstream tasks. For example, combining diffusion and hyperbolic geometry can be extended to deep-based networks (Ganea et al., 2018; Chami et al., 2019), which have been shown useful for hierarchical data learning both theoretically and empirically.

Acknowledgements

We express our gratitude to the anonymous reviewers for their valuable feedback. The work of YEL and RT was supported by the European Union's Horizon 2020 research and innovation programme under grant agreement No. 802735-ERC-DIFFOP. The work of GM was supported by NSF award CCF-2217058.

References

Beardon, A. F. *The geometry of discrete groups*, volume 91. Springer Science & Business Media, 2012.

Belkin, M. and Niyogi, P. Towards a theoretical foundation for Laplacian-based manifold methods. *Journal of Computer and System Sciences*, 74(8):1289–1308, 2008.

Bronstein, A. M., Bronstein, M. M., Kimmel, R., Mahmoudi, M., and Sapiro, G. A Gromov-Hausdorff framework with diffusion geometry for topologically-robust non-rigid shape matching. *International Journal of Computer Vision*, 89(2):266–286, 2010.

Cetin, E., Chamberlain, B., Bronstein, M., and Hunt, J. J. Hyperbolic deep reinforcement learning. *arXiv preprint arXiv:2210.01542*, 2022.

Chamberlain, B. P., Clough, J., and Deisenroth, M. P. Neural embeddings of graphs in hyperbolic space. *arXiv preprint arXiv:1705.10359*, 2017.

Chami, I., Ying, Z., Ré, C., and Leskovec, J. Hyperbolic graph convolutional neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

Chami, I., Gu, A., Chatziafratis, V., and Ré, C. From trees to continuous embeddings and back: Hyperbolic hierarchical clustering. *Advances in Neural Information Processing Systems*, 33:15065–15076, 2020.

Cheng, X., Rachh, M., and Steinerberger, S. On the diffusion geometry of graph Laplacians and applications. *Applied and Computational Harmonic Analysis*, 46(3): 674–688, 2019.

Coifman, R. R. and Goldberg, M. J. Some extensions of E. Stein's work on Littlewood–Paley theory applied to symmetric diffusion semigroups. *The Journal of Geometric Analysis*, 31(7):6781–6795, 2021.

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.

Cox, M. A. and Cox, T. F. Multidimensional scaling. In *Handbook of data visualization*, pp. 315–347. Springer, 2008.

Dasgupta, S. A cost function for similarity-based hierarchical clustering. In *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing*, pp. 118–127, 2016.

De Nooy, W., Mrvar, A., and Batagelj, V. *Exploratory social network analysis with Pajek: Revised and expanded edition for updated software*, volume 46. Cambridge university press, 2018.

Ding, J. and Regev, A. Deep generative model embedding of single-cell RNA-Seq profiles on hyperspheres and hyperbolic spaces. *Nature communications*, 12(1):1–17, 2021.

Ding, X. and Wu, H.-T. Impact of signal-to-noise ratio and bandwidth on graph Laplacian spectrum from high-dimensional noisy point cloud. *arXiv preprint arXiv:2011.10725*, 2020.

Dua, D. and Graff, C. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Dumitrascu, B., Villar, S., Mixon, D. G., and Engelhardt, B. E. Optimal marker gene selection for cell type discrimination in single cell analyses. *Nature communications*, 12(1):1–8, 2021.

duVerle, D. A., Yotsukura, S., Nomura, S., Aburatani, H., and Tsuda, K. CellTree: an R/bioconductor package to infer the hierarchical structure of cell populations from single-cell RNA-Seq data. *BMC bioinformatics*, 17(1): 1–17, 2016.

Fang, P., Harandi, M., and Petersson, L. Kernel methods in hyperbolic spaces. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 10665–10674, 2021.

Fefferman, C., Mitter, S., and Narayanan, H. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Ficken, F. A. The Riemannian and affine differential geometry of product-spaces. *Annals of Mathematics*, pp. 892–913, 1939.

Frank, R. L. and Kovarik, H. Heat kernels of metric trees and applications. *SIAM Journal on Mathematical Analysis*, 45(3):1027–1046, 2013.

Ganea, O., Bécigneul, G., and Hofmann, T. Hyperbolic neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.

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 *ICML*, 2010.

Goldberg, M. J. and Kim, S. An efficient tree-based computation of a metric comparable to a natural diffusion distance. *Applied and Computational Harmonic Analysis*, 33(2):261–281, 2012.

Grigoryan, A. *Heat kernel and analysis on manifolds*, volume 47. American Mathematical Soc., 2009.

Grigor'yan, A. and Noguchi, M. The heat kernel on hyperbolic space. *Bulletin of the London Mathematical Society*, 30(6):643–650, 1998.

Gromov, M. Hyperbolic groups. In *Essays in group theory*, pp. 75–263. Springer, 1987.

Gu, A., Sala, F., Gunel, B., and Ré, C. Learning mixed-curvature representations in product spaces. In *International Conference on Learning Representations*, 2018.

Hellinger, E. Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 1909(136):210–271, 1909.

Hofbauer, W. K., Forrest, L. L., Hollingsworth, P. M., and Hart, M. L. Preliminary insights from DNA barcoding into the diversity of mosses colonising modern building surfaces. *Bryophyte Diversity and Evolution*, 38(1):1–22, 2016.

Jaskowiak, P. A., Campello, R. J., and Costa, I. G. On the selection of appropriate distances for gene expression data clustering. In *BMC bioinformatics*, volume 15, pp. 1–17. Springer, 2014.

Jordan, M. I. and Mitchell, T. M. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

Keller, Y., Coifman, R. R., Lafon, S., and Zucker, S. W. Audio-visual group recognition using diffusion maps. *IEEE Transactions on Signal Processing*, 58(1):403–413, 2009.

Khrulkov, V., Mirvakhabova, L., Ustinova, E., Oseledets, I., and Lempitsky, V. Hyperbolic image embeddings. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6418–6428, 2020.

Klimovskaia, A., Lopez-Paz, D., Bottou, L., and Nickel, M. Poincaré maps for analyzing complex hierarchies in single-cell data. *Nature communications*, 11(1):1–9, 2020.

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(4):910–953, 2016.

Leeb, W. E. *Topics in metric approximation*. 2015.

Leskovec, J., Kleinberg, J., and Faloutsos, C. Graph evolution: Densification and shrinking diameters. *ACM transactions on Knowledge Discovery from Data (TKDD)*, 1 (1):2–es, 2007.

Li, H., Lindenbaum, O., Cheng, X., and Cloninger, A. Variational diffusion autoencoders with random walk sampling. In *European Conference on Computer Vision*, pp. 362–378. Springer, 2020.

Lin, Y.-W. E., Kluger, Y., and Talmon, R. Hyperbolic procrustes analysis using Riemannian geometry. *Advances in Neural Information Processing Systems*, 34:5959–5971, 2021.

Lindenbaum, O., Yeredor, A., Salhov, M., and Averbuch, A. Multi-view diffusion maps. *Information Fusion*, 55: 127–149, 2020.

McKean, H. P. An upper bound to the spectrum of Δ on a manifold of negative curvature. *Journal of Differential Geometry*, 4(3):359–366, 1970.

Mishne, G. and Cohen, I. Multiscale anomaly detection using diffusion maps. *IEEE Journal of selected topics in signal processing*, 7(1):111–123, 2012.

Mishne, G., Talmon, R., Meir, R., Schiller, J., Lavzin, M., Dubin, U., and Coifman, R. R. Hierarchical coupled-geometry analysis for neuronal structure and activity pattern discovery. *IEEE Journal of Selected Topics in Signal Processing*, 10(7):1238–1253, 2016.

Mishne, G., Talmon, R., Cohen, I., Coifman, R. R., and Kluger, Y. Data-driven tree transforms and metrics. *IEEE transactions on signal and information processing over networks*, 4(3):451–466, 2017.

Mishne, G., Chi, E., and Coifman, R. Co-manifold learning with missing data. In *International Conference on Machine Learning*, pp. 4605–4614. PMLR, 2019.

Mishne, G., Wan, Z., Wang, Y., and Yang, S. The numerical stability of hyperbolic representation learning. *arXiv preprint arXiv:2211.00181*, 2022.

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

Murphy, J. M. and Maggioni, M. Spectral–spatial diffusion geometry for hyperspectral image clustering. *IEEE Geoscience and Remote Sensing Letters*, 17(7):1243–1247, 2019.

Nadler, B., Lafon, S., Kevrekidis, I., and Coifman, R. Diffusion maps, spectral clustering and eigenfunctions of Fokker-Planck operators. *Advances in Neural Information Processing Systems*, 18, 2005.

Nickel, M. and Kiela, D. Poincaré embeddings for learning hierarchical representations. *Advances in Neural Information Processing Systems*, 30, 2017.

Nickel, M. and Kiela, D. Learning continuous hierarchies in the Lorentz model of hyperbolic geometry. In *International Conference on Machine Learning*, pp. 3779–3788. PMLR, 2018.

Ollivier, Y. A visual introduction to Riemannian curvatures and some discrete generalizations. *Analysis and Geometry of Metric Measure Spaces: Lecture Notes of the 50th Séminaire de Mathématiques Supérieures (SMS), Montréal*, 56:197–219, 2011.

Sala, F., De Sa, C., Gu, A., and Ré, C. Representation tradeoffs for hyperbolic embeddings. In *International conference on machine learning*, pp. 4460–4469. PMLR, 2018.

Sanderson, M., Donoghue, M., Piel, W., and Eriksson, T. Treebase: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. *American Journal of Botany*, 81(6):183, 1994.

Sarkar, R. Low distortion delaunay embedding of trees in hyperbolic plane. In *International Symposium on Graph Drawing*, pp. 355–366. Springer, 2011.

Shen, C. and Wu, H.-T. Scalability and robustness of spectral embedding: landmark diffusion is all you need. *Information and Inference: A Journal of the IMA*, 11(4): 1527–1595, 2022.

Shnitzer, T., Yurochkin, M., Greenewald, K., and Solomon, J. M. Log-Euclidean signatures for intrinsic distances between unaligned datasets. In *International Conference on Machine Learning*, pp. 20106–20124. PMLR, 2022.

Singer, A. From graph to manifold Laplacian: The convergence rate. *Applied and Computational Harmonic Analysis*, 21(1):128–134, 2006.

Skopek, O., Ganea, O.-E., and Bécigneul, G. Mixed-curvature variational autoencoders. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=S1g6xeSKDS.

Sonthalia, R. and Gilbert, A. Tree! I am no Tree! I am a low dimensional hyperbolic embedding. *Advances in Neural Information Processing Systems*, 33:845–856, 2020.

Stoeckius, M., Hafemeister, C., Stephenson, W., Houck-Loomis, B., Chattopadhyay, P. K., Swerdlow, H., Satija, R., and Smibert, P. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 14 (9):865–868, 2017.

Talmon, R. and Coifman, R. R. Empirical intrinsic geometry for nonlinear modeling and time series filtering. *Proceedings of the National Academy of Sciences*, 110 (31):12535–12540, 2013.

Talmon, R., Cohen, I., Gannot, S., and Coifman, R. R. Diffusion maps for signal processing: A deeper look at manifold-learning techniques based on kernels and graphs. *IEEE signal processing magazine*, 30(4):75–86, 2013.

Tanay, A. and Regev, A. Scaling single-cell genomics from phenomenology to mechanism. *Nature*, 541(7637):331–338, 2017.

Tay, Y., Tuan, L. A., and Hui, S. C. Hyperbolic representation learning for fast and efficient neural question answering. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining*, pp. 583–591, 2018.

Tifrea, A., Bécigneul, G., and Ganea, O.-E. Poincaré glove: Hyperbolic word embeddings. *arXiv preprint arXiv:1810.06546*, 2018.

Tong, A. Y., Huguet, G., Natik, A., MacDonald, K., Kuchroo, M., Coifman, R., Wolf, G., and Krishnaswamy, S. Diffusion earth mover's distance and distribution embeddings. In *International Conference on Machine Learning*, pp. 10336–10346. PMLR, 2021.

Turaga, P. K. and Srivastava, A. *Riemannian computing in computer vision*, volume 1. Springer, 2016.

van Galen, P., Hovestadt, V., Wadsworth II, M. H., Hughes, T. K., Griffin, G. K., Battaglia, S., Verga, J. A., Stephansky, J., Pastika, T. J., Story, J. L., et al. Single-cell RNA-seq reveals AML hierarchies relevant to disease progression and immunity. *Cell*, 176(6):1265–1281, 2019.

Verbeek, K. and Suri, S. Metric embedding, hyperbolic space, and social networks. In *Proceedings of the thirtieth annual symposium on Computational geometry*, pp. 501–510, 2014.

Zeisel, A., Muñoz-Manchado, A. B., Codeluppi, S., Lönnerberg, P., La Manno, G., Juréus, A., Marques, S., Munguba, H., He, L., Betsholtz, C., et al. Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq. *Science*, 347(6226):1138–1142, 2015.

Zelditch, S. *Eigenfunctions of the Laplacian on a Riemannian manifold*, volume 125. American Mathematical Soc., 2017.

Zelnik-Manor, L. and Perona, P. Self-tuning spectral clustering. *Advances in Neural Information Processing Systems*, 17, 2004.

A. Additional Background

A.1. Diffusion Geometry

Broadly, the construction of diffusion geometry starts by defining a probability transition matrix $\mathbf{P}$ that describes how likely it is to transition from one data point to another. This matrix is then used to construct a Markov process, which defines the diffusion distance between data points, conveying a notion of distance between data points based on how easily one can transition or "diffuse" from one point to another. Formally, the diffusion distance with time diffusion t between two points $\mathbf{x}$ and $\mathbf{x}'$ is given by $\|\mathbf{P}^t e_i - \mathbf{P}^t e_{i'}\|$ with an appropriate norm (see (Coifman & Lafon, 2006)). This diffusion distance is then used to construct a family of multi-scale low-dimensional maps of the data set, termed diffusion maps. The diffusion maps in $\ell < n$ dimensions with diffusion time t of a point $\mathbf{x}$ is given by $\Psi_t(\mathbf{x}_i) = [\lambda_1^t \nu_1(i), \ldots, \lambda_\ell^t \nu_\ell(i)]^\top$, where $\{(\lambda_i, \nu_i)\}_{i=1}^n$ are the eigen-pairs of the transition matrix $\mathbf{P}$. It is shown that the Euclidean distances between the diffusion maps approximate the diffusion distances (Coifman & Lafon, 2006). In (Coifman & Maggioni, 2006), diffusion operators with multiple scales t on a dyadic grid were considered for multi-scale data representation, called diffusion wavelets.

A.2. Hyperbolic Geometry

The n-dimensional Poincaré half-space (Beardon, 2012) is a Riemannian manifold with constant negative curvature, defined by $\mathbb{H}^n = \{\mathbf{x} \in \mathbb{R}^n \big| \mathbf{x}(n) > 0\}$ with the Riemannian metric tensor $ds^2 = \frac{d\mathbf{x}^2(1) + d\mathbf{x}^2(2) + \ldots + d\mathbf{x}^2(n)}{a^2 \mathbf{x}^2(n)}$, where $a > 0$ and $\kappa = -a^2$ represents the Gaussian curvature of the hyperbolic manifold. In this work, we study the n-dimensional Poincaré half-space with constant negative curvature -1 by setting $a = 1$.

The hyperbolic geometry can be characterized by Gromov's δ-hyperbolicity (Gromov, 1987; Ollivier, 2011).

Definition A.1. *A metric space (X, d) is δ-hyperbolic (Gromov, 1987) if there exists $\delta \geq 0$ such that for all four points $x, y, z, w \in X$*

$$d(w, x) + d(y, z) \leq \max\{d(x, y) + d(w, z), d(x, z) + d(y, w)\} + 2\delta. \tag{15}$$

Proposition A.1. *A 0-hyperbolic metric satisfies the triangle inequality:*

$$d(w, x) \leq d(w, y) + d(x, y). \tag{16}$$

Example A.1. *The two-dimensional Poincare half-plane $\mathbb{H}^2$ is $(\log 2)$-hyperbolic.*

A.3. Graph Preliminaries

Definition A.2. *Let $G = (V, E, W)$ be an undirected graph. The shortest path metric $d_{\mathcal{T}}(u, v)$ is the length of the shortest path from u to v.*

Definition A.3. *A metric d is a 0-hyperbolic metric if there exists a tree $\mathcal{T}$ such that the shortest path metric $d_{\mathcal{T}}$ on $\mathcal{T}$ is equal to d.*

B. Proof of Proposition 1

Proposition 1. There is a constant $0 < C < 1$ such that for any $x_i, x_{i'} \in \mathcal{X}$ and $k_1 \leq k_2$ for $k_1, k_2 \in \mathbb{Z}_0^+$, we have

$$C \cdot 2^{-(k_2 - k_1)\alpha} \leq \frac{d_{\mathbb{H}^{n+1}}(\hat{\mathbf{x}}_i^{k_2}, \hat{\mathbf{x}}_{i'}^{k_2})}{d_{\mathbb{H}^{n+1}}(\hat{\mathbf{x}}_i^{k_1}, \hat{\mathbf{x}}_{i'}^{k_1})} \leq \frac{1}{C} \cdot 2^{-(k_2 - k_1)\alpha}. \tag{17}$$

Proof. For any $k_1, k_2 \in \mathbb{Z}_0^+$ such that $k_1 \leq k_2$ and $x_i, x_{i'} \in \mathcal{X}$, by the bounds of the Hellinger distance, we have $0 < c \leq \left\|\varphi_i^{k_1} - \varphi_{i'}^{k_1}\right\|_2 \leq \left\|\varphi_i^{k_2} - \varphi_{i'}^{k_2}\right\|_2 \leq \sqrt{2}$ for some constant c. We begin with the proof of lower-bound:

$$\begin{aligned}
\frac{d_{\mathbb{H}^{n+1}}(\hat{\mathbf{x}}_i^{k_2}, \hat{\mathbf{x}}_{i'}^{k_2})}{d_{\mathbb{H}^{n+1}}(\hat{\mathbf{x}}_i^{k_1}, \hat{\mathbf{x}}_{i'}^{k_1})} &= \frac{\sinh^{-1}\left(2^{-k_2\alpha+1}\left\|\varphi_i^{k_2} - \varphi_{i'}^{k_2}\right\|_2\right)}{\sinh^{-1}\left(2^{-k_1\alpha+1}\left\|\varphi_i^{k_1} - \varphi_{i'}^{k_1}\right\|_2\right)} \\
&\geq \frac{\sinh^{-1}\left(2^{-k_2\alpha+1} \cdot c\right)}{\sinh^{-1}\left(2^{-k_1\alpha+1} \cdot \sqrt{2}\right)}
\end{aligned}$$

$$\overset{(1)}{\geq} \frac{\frac{1}{2}\left(2^{-k_2\alpha+1}\cdot c\right)}{\left(2^{-k_1\alpha+1}\cdot\sqrt{2}\right)} = \frac{c}{2\sqrt{2}}\cdot 2^{-(k_2-k_1)\alpha},$$

where transition (1) is due to $\sinh^{-1}(z) < z$ for $z > 0$ and $2\sinh^{-1}(2^{-k\alpha+1}z) > 2^{-k\alpha+1}z$ for $0 < z < \sqrt{2}$, $0 < \alpha < 1$, and $k \in \mathbb{Z}_0^+$. Similarly, the upper bound is obtained by

$$\begin{aligned}\frac{d_{\mathbb{H}^{n+1}}(\hat{\mathbf{x}}_i^{k_2},\hat{\mathbf{x}}_{i'}^{k_2})}{d_{\mathbb{H}^{n+1}}(\hat{\mathbf{x}}_i^{k_1},\hat{\mathbf{x}}_{i'}^{k_1})} &= \frac{\sinh^{-1}\left(2^{-k_2\alpha+1}\left\|\varphi_i^{k_2}-\varphi_{i'}^{k_2}\right\|_2\right)}{\sinh^{-1}\left(2^{-k_1\alpha+1}\left\|\varphi_i^{k_1}-\varphi_{i'}^{k_1}\right\|_2\right)}\\ &\leq \frac{\sinh^{-1}\left(2^{-k_2\alpha+1}\cdot\sqrt{2}\right)}{\sinh^{-1}\left(2^{-k_1\alpha+1}\cdot c\right)}\\ &\leq \frac{2^{-k_2\alpha+1}\cdot\sqrt{2}}{\frac{1}{2}\cdot 2^{-k_1\alpha+1}\cdot c}\\ &= \frac{2\sqrt{2}}{c}\cdot 2^{-(k_2-k_1)\alpha}.\end{aligned}$$

Taking $C = \frac{c}{2\sqrt{2}}$ gives the results. We remark that the lower bound can be tightly bounded by $2^{-(k_2-k_1)\alpha}$ due to $\frac{d_{\mathbb{H}^{n+1}}(\hat{\mathbf{x}}_i^{k_2},\hat{\mathbf{x}}_{i'}^{k_2})}{d_{\mathbb{H}^{n+1}}(\hat{\mathbf{x}}_i^{k_1},\hat{\mathbf{x}}_{i'}^{k_1})} = \frac{\sinh^{-1}\left(2^{-k_2\alpha+1}\left\|\varphi_i^{k_2}-\varphi_{i'}^{k_2}\right\|_2\right)}{\sinh^{-1}\left(2^{-k_1\alpha+1}\left\|\varphi_i^{k_1}-\varphi_{i'}^{k_1}\right\|_2\right)} \geq \frac{\sinh^{-1}\left(2^{-k_2\alpha+1}\left\|\varphi_i^{k_2}-\varphi_{i'}^{k_2}\right\|_2\right)}{\sinh^{-1}\left(2^{-k_1\alpha+1}\left\|\varphi_i^{k_2}-\varphi_{i'}^{k_2}\right\|_2\right)} = 2^{-(k_2-k_1)\alpha}$. □

C. Theoretical Analysis of Hyperbolic Diffusion Distance - Proof of Theorem 1

The theoretical analysis of HDD is motivated by and derived from the work presented in (Leeb & Coifman, 2016). In their work, the authors considered the geometric regularity conditions on the diffusion semi-group of a multi-scale total variation distance between probability measures (Goldberg & Kim, 2012; Leeb, 2015). More specifically, they presented a diffusion ground distance, a multi-scale distance using L_1 distance between probability measures for approximating the geodesic distance on a closed Riemannian manifold with non-negative curvature.

In our work, we focus on the hierarchical (i.e., tree or tree-like) structures that cannot be approximated by the work in (Leeb & Coifman, 2016). To this end, we follow the work of (McKean, 1970; Grigor'yan & Noguchi, 1998; Frank & Kovarik, 2013; Zelditch, 2017) for spaces with negative curvature and devise the HDD based on a multi-scale metric using inverse hyperbolic sine function of a scaled Hellinger distance (Hellinger, 1909), which forms the ℓ_1 distance on the product manifold of the hyperbolic spaces.

First, we define the multi-scale metric HDD in a continuous space and introduce the properties of the diffusion semi-group. Next, we establish the geometric regularities in the case of hierarchical datasets that are necessary for the multi-scale metric to approximate the underlying tree metric. Last, we will show that the diffusion operators, which approximate the heat kernel, satisfy these conditions, and therefore, the proposed HDD recovers the hierarchical structure underlying the data.

C.1. HDD in Continuous Space

Let $\mathcal{X}$ be a sigma-finite measure space in dimension n. We consider a measure μ such that $\mu(B(x,r)) \lesssim r^n$, where $x \in \mathcal{X}$ and $r > 0$. A family of kernels $\{a_t(x,x')\}_{t\in\mathbb{R}_+}$ is considered for $x, x' \in \mathcal{X}$. Let f be a function defined in $\mathcal{X}$. We define the operator A_t by

$$A_t f(x) = \int_{\mathcal{X}} a_t(x,x')f(x')dx'. \tag{18}$$

The operators $\{A_t\}_{t\in\mathbb{R}_+}$ have the following properties (Coifman & Lafon, 2006; Coifman & Goldberg, 2021). (i) The family of operators forms a semi-group such that for all $t_1, t_2 \in \mathbb{R}_+$, we have $A_{t_1}A_{t_2} = A_{t_1+t_2}$. (ii) The operator respects the conservation property, that is, $\int_{\mathcal{X}} a_t(x,x')dx' = 1$. (iii) The operator is integrable such that $\int_{\mathcal{X}} |a_t(x,x')|dy \leq C$ for some constant $C > 0$.

We are only concerned with dyadic times $t \in (0,1]$ such that $t = 2^{-k}$ for $k \in \mathbb{Z}_0^+$. We first define the local geometric measure at a single scale k using the unnormalized Hellinger distance (Hellinger, 1909) between the probability distributions, given by

$$T_k(x,x') = \left\| \sqrt{a_{2^{-k}}(x,\cdot)} - \sqrt{a_{2^{-k}}(x',\cdot)} \right\|_2 . \tag{19}$$

Then we define the multi-scale metric using the inverse hyperbolic sine function of the scaled Hellinger measure with the scaling term $2^{-k\alpha+1}$, given by

$$\hat{T}_\alpha(x,x') = \sum_{k\geq 0} 2\sinh^{-1}\left(2^{-k\alpha+1}\, T_k(x,x')\right) = \sum_{k\geq 0} 2\sinh^{-1}\left(2^{-k\alpha+1} \left\| \sqrt{a_{2^{-k}}(x,\cdot)} - \sqrt{a_{2^{-k}}(x',\cdot)} \right\|_2 \right), \tag{20}$$

where $0 < \alpha < 1$. Because the scaling parameters decay exponentially, the multi-scale metric can be approximated by the first K terms:

$$\hat{T}_\alpha(x,x') \approx \hat{T}_K(x,x') = \sum_{k=0}^{K} 2\sinh^{-1}\left(2^{-k\alpha+1}\, T_k(x,x')\right). \tag{21}$$

C.2. Regularity Conditions

We impose geometric regularity on the multi-scale metric $\hat{T}_\alpha$. There are constants $C > 0$ and $\alpha > 0$ such that the integral of the kernel and the multi-scale metric at scale k is upper-bounded by

$$\int_{\mathcal{X}} a_{2^{-k}}(x,x')\hat{T}_\alpha(x,x')dx' \leq C2^{-k\alpha}. \tag{22}$$

Let $(\mathcal{X}, d_\mathcal{T})$ be a hierarchical metric space. There are three strong regularity conditions imposed on the family of operators $\{A_{2^{-k}}\}_{k\in\mathbb{R}_+}$ that allow for the proposed multi-metric $\hat{T}_\alpha^K(x,x')$ to approximate $d_\mathcal{T}$.

The first condition is *an upper bound on the kernel*. There is a non-negative and monotonic decreasing function $f_1 : \mathbb{R}_+ \to \mathbb{R}$ and a number $\beta > 0$ such that for any $\gamma < \beta$, we have $\int_{\mathbb{R}_+} \tau^{3n+\gamma} f_1(\tau)d\tau/\tau < \infty$. The square root of the kernel for all $t \in (0,1]$ is then upper-bounded by

$$\sqrt{a_t(x,x')} \leq \frac{1}{t^{\frac{3n}{2\beta}}} f_1\left(\frac{d_\mathcal{T}(x,x')}{t^{\frac{1}{\beta}}}\right). \tag{23}$$

The second condition is *a lower bound of the kernel*. There is a monotonic decreasing function $g_1 : \mathbb{R}_+ \to \mathbb{R}$ and $R > 0$ such that for all $t \in (0,1]$ and all $d_\mathcal{T}(x,x') < R$, the square root of the kernel is lower-bounded by

$$\sqrt{a_t(x,x')} \geq \frac{1}{t^{\frac{n}{2\beta}}} g_1\left(\frac{d_\mathcal{T}(x,x')}{t^{\frac{1}{\beta}}}\right). \tag{24}$$

The third condition is *Hölder continuity*. There is a constant $\Theta > 0$ sufficiently small such that for all $t \in (0,1]$, all $x, x' \in \mathcal{X}$ with $d_\mathcal{T}(x,x') \leq t^{\frac{1}{\beta}}$ and all $y \in \mathcal{X}$, the element value of the Hellinger measure is upper-bounded by

$$|\sqrt{a_t(x,y)} - \sqrt{a_t(x',y)}|^2 \leq \left(\frac{d_\mathcal{T}(x,x')}{t^{\frac{1}{\beta}}}\right)^{2\Theta} \frac{1}{t^{\frac{n}{\beta}}} f_1\left(\frac{d_\mathcal{T}(x,y)}{t^{\frac{1}{\beta}}}\right). \tag{25}$$

C.3. Hierarchical Metric

We present the lower and upper bounds of the proposed multi-scale metric $\hat{T}_\alpha$ in Eq. (20), making it equivalent to the hierarchical metric $d_\mathcal{T}$.

Definition C.1 (Snowflake distance (Leeb, 2015; Leeb & Coifman, 2016))**.** *The snowflake distance is a distance in the form of $d(\cdot,\cdot)^s$, where d is a distance and $0 < s < 1$.*

We first present the upper bound of $\hat{T}_\alpha$.

Proposition C.1. *For any $0 < \alpha < \min\{1, \frac{\Theta}{\beta}\}$, the multi-scale $\hat{T}_\alpha$ is upper-bounded by*

$$\hat{T}_\alpha(x,x') \lesssim \min\{1, d_\mathcal{T}^{\alpha\beta}(x,x')\}. \tag{26}$$

Proof. Consider the dyadic levels $K-1$ and K such that the tree distance is bounded from below and above by $2^{-K} \lesssim d_{\mathcal{T}}^{\beta}(x,x') \lesssim 2^{-K+1}$. We have

$$\begin{aligned}
\hat{T}_{\alpha}(x,x') &= \sum_{k\geq 0} 2\sinh^{-1}\left(2^{-k\alpha+1}\left\|\sqrt{a_{2^{-k}}(x,\cdot)}-\sqrt{a_{2^{-k}}(x',\cdot)}\right\|_2\right) \\
&\overset{(1)}{\leq} \sum_{k\geq 0} 2\cdot 2^{-k\alpha+1}\left\|\sqrt{a_{2^{-k}}(x,\cdot)}-\sqrt{a_{2^{-k}}(x',\cdot)}\right\|_2 \\
&= \sum_{k\geq 0} 2^{-k\alpha+2}\left\|\sqrt{a_{2^{-k}}(x,\cdot)}-\sqrt{a_{2^{-k}}(x',\cdot)}\right\|_2 \\
&\overset{(2)}{\lesssim} d_{\mathcal{T}}^{\Theta}(x,x')\sum_{k=0}^{K} 2^{-k\alpha+2}2^{\frac{k\Theta}{\beta}} + \sum_{k=K+1}^{\infty} 2^{-k\alpha+2} \\
&\lesssim d_{\mathcal{T}}^{\Theta}(x,x')2^{K\alpha}2^{\frac{K\Theta}{\beta}} + 2^{-K\alpha} \\
&\overset{(3)}{\lesssim} d_{\mathcal{T}}^{\alpha\beta}(x,x'),
\end{aligned}$$

where transition (1) is due to $\sinh^{-1}(z) < z$ for $z > 0$, transition (2) is based on the Hölder continuity condition in Eq. (25) implying that the Hellinger distance is bounded by $\left\|\sqrt{a_{2^{-k}}(x,\cdot)}-\sqrt{a_{2^{-k}}(x',\cdot)}\right\|_2 \lesssim \left(\frac{d_{\mathcal{T}}(x,x')}{t^{\frac{1}{\beta}}}\right)^{\Theta}$, and transition (3) is due to $\alpha < \frac{\Theta}{\beta}$. □

Proposition C.1 implies that the upper-bound of the multi-scale metric $\hat{T}_{\alpha}$ is a thresholded Snowflake distance of $d_{\mathcal{T}}$. Below, we will demonstrate the lower bound using the following two results.

Lemma C.1. *Let p, q be two probability distributions on $\mathcal{X}$. For a constant $k \in \mathbb{Z}_0^+$ and $0 < \alpha < 1$, we have*

$$2\sinh^{-1}\left(2^{-k\alpha+1}\left\|\sqrt{p}-\sqrt{q}\right\|_2\right) \geq 2^{-k\alpha}\|p-q\|_1, \tag{27}$$

where $\left\|\sqrt{p}-\sqrt{q}\right\|_2$ is the unnormalized Hellinger distance between p and q.

Next, we introduce the lower bound of $\hat{T}_{\alpha}$.

Lemma C.2 (Lemma 3 in (Leeb & Coifman, 2016))**.** *Let R be the condition of the lower bound of the kernel in Eq. (24). There are constants $A > 1$ and $\epsilon > 0$ such that whenever $x, x' \in \mathcal{X}$ and $t \in (0,1]$ satisfy $At^{\frac{1}{\beta}} \leq d_{\mathcal{T}}(x,x') < R$, we have*

$$\|a_{2^{-k}}(x,\cdot) - a_{2^{-k}}(x',\cdot)\|_1 \geq \epsilon. \tag{28}$$

Proposition C.2. *Let R be as in the condition of the lower bound of the kernel in Eq. (24). The multi-scale metric is lower-bounded by*

$$\hat{T}_{\alpha}(x,x') \gtrsim d_{\mathcal{T}}^{\alpha\beta}(x,x'). \tag{29}$$

Proof. Take A and ϵ as in the conditions in Lemma C.2. We now take $K \in \mathbb{Z}_+$ such that $\frac{d_{\mathcal{T}}^{\beta}(x,x)}{A^{\beta}}$ is bounded by $2^{-K} \lesssim \frac{d_{\mathcal{T}}^{\beta}(x,x')}{A^{\beta}} \lesssim 2^{-K+1}$. Then, we have

$$\begin{aligned}
\hat{T}_{\alpha}(x,x') &= \sum_{k\geq 0} 2\sinh^{-1}\left(2^{-k\alpha+1}\left\|\sqrt{a_{2^{-k}}(x,\cdot)}-\sqrt{a_{2^{-k}}(x',\cdot)}\right\|_2\right) \\
&\geq \sum_{k\geq K} 2\sinh^{-1}\left(2^{-k\alpha+1}\left\|\sqrt{a_{2^{-k}}(x,\cdot)}-\sqrt{a_{2^{-k}}(x',\cdot)}\right\|_2\right) \\
&\overset{(1)}{\geq} \sum_{k\geq K} 2^{-k\alpha}\|a_{2^{-k}}(x,\cdot) - a_{2^{-k}}(x',\cdot)\|_1 \\
&\overset{(2)}{\geq} \epsilon \sum_{k\geq K} 2^{-k\alpha}
\end{aligned}$$

$$\simeq 2^{-K\alpha}$$
$$\simeq d_{\mathcal{T}}^{\alpha\beta}(x,x')$$

where transition (1) is based on Lemma C.1 and transition (2) is applied by Lemma C.2 using the triangle inequality in Eq. (16). □

Lemma C.3 (Lemma 4 in (Leeb & Coifman, 2016)). *Let R be the condition of the lower bound of the kernel in Eq. (24). There are constants $C > 0$ and $\eta > 0$ such that whenever $d_{\mathcal{T}}(x,x') \geq R$ and $t^{\frac{1}{\beta}} < \eta R$, we have*

$$\|a_{2^{-k}}(x,\cdot) - a_{2^{-k}}(x',\cdot)\|_1 \geq C. \tag{30}$$

Proposition C.3. *Let R be the condition of the lower bound of the kernel in Eq. (24). There is a constant $C > 0$ such that when $d_{\mathcal{T}}(x,x') > R$, we have*

$$\hat{T}_\alpha(x,x') \gtrsim C. \tag{31}$$

Proof. Take C and η as in the conditions in Lemma C.3. Let $K = \left\lfloor \log_2\left(\frac{1}{(\eta R)^\beta}\right)\right\rfloor$. Then $2^{-K} \leq (\eta R)^\beta$, and we have

$$\begin{aligned}
\hat{T}_\alpha(x,x') &= \sum_{k\geq 0} 2\sinh^{-1}\left(2^{-k\alpha+1}\left\|\sqrt{a_{2^{-k}}(x,\cdot)} - \sqrt{a_{2^{-k}}(x',\cdot)}\right\|_2\right) \\
&\geq \sum_{k\geq K} 2\sinh^{-1}\left(2^{-k\alpha+1}\left\|\sqrt{a_{2^{-k}}(x,\cdot)} - \sqrt{a_{2^{-k}}(x',\cdot)}\right\|_2\right) \\
&\overset{(1)}{\geq} \sum_{k\geq K} 2^{-k\alpha}\|a_{2^{-k}}(x,\cdot) - a_{2^{-k}}(x',\cdot)\|_1 \\
&\overset{(2)}{\geq} \sum_{k\geq K} 2^{-k\alpha} C \\
&\simeq C(\eta R)^{\alpha\beta},
\end{aligned}$$

where transition (1) is derived by Lemma C.1 and transition (2) is based on Lemma C.3. □

Proposition C.2 and Proposition C.3 guarantee that the lower-bound of the multi-scale metric is a thresholded Snowflake distance of $d_{\mathcal{T}}$. We summarize it in the following corollary.

Corollary C.1. *Under the conditions of upper and lower bounds of the kernel and the Hölder continuity in Eq. (23), Eq. (24), and Eq. (25), we have*

$$\hat{T}_\alpha(x,x') \gtrsim \min\{1, d_{\mathcal{T}}^{\alpha\beta}(x,x')\}. \tag{32}$$

Last, we summarize the equivalence of $\hat{T}_\alpha$ to a thresholded snowflake metric by using Proposition C.1 and Corollary C.1.

Proposition C.4. *If the conditions for the upper and lower bounds of the kernel and the Hölder continuity on $a_{2^{-k}}(x,x')$ hold and if $\mu(B(x,r)) \lesssim r^n$, then for $0 < \alpha < \min\{1, \frac{\Theta}{\beta}\}$ the distance $\hat{T}_\alpha(x,x')$ is equivalent to the thresholded snowflake distance $\min\{1, d_{\mathcal{T}}^{\alpha\beta}(x,x')\}$.*

C.4. Heat Kernel on Trees

In the following, we show that Proposition C.4 holds for the *heat kernel on a tree*. For this purpose, we follow (McKean, 1970; Grigor'yan & Noguchi, 1998; Frank & Kovarik, 2013; Zelditch, 2017), showing that the necessary conditions imposed on $\{a_t(x,x')\}_{t\in\mathbb{R}_+}$ are satisfied.

In the following lemmas, the operator a_t is considered as the heat kernel on tree.

Lemma C.4. *There are constants $A, B > 0$ such that for all $t \in (0,1]$ we have*

$$a_t(x,x') \leq \frac{A}{t^{\frac{3n}{2}}} \exp\left(-\frac{B \cdot d_{\mathcal{T}}^2(x,x')}{t}\right). \tag{33}$$

Lemma C.5. *There are constants $C, D > 0$ such that for all $t \in (0, 1]$ we have*

$$a_t(x, x') \geq \frac{C}{t^{\frac{n}{2}}} \exp\left(-\frac{D \cdot d_{\mathcal{T}}^2(x, x')}{t}\right). \tag{34}$$

Lemma C.6. *There are constants $E, F > 0$ such that for $t \in (0, 1]$ we have*

$$\|\nabla_x a_t(x, x')\|_2 \leq \frac{E}{\sqrt{t} \cdot t^{\frac{n}{2}}} \exp\left(-\frac{F \cdot d_{\mathcal{T}}^2(x, x')}{t}\right). \tag{35}$$

Proposition C.5. *If $x, x' \in \mathcal{X}$ are sufficiently close, then for a smooth function $h : \mathcal{X} \to \mathbb{R}_+$, there is a point y lying on the path on $\mathcal{T}$ from x to y such that*

$$|h^{\frac{1}{2}}(x) - h^{\frac{1}{2}}(x')|^2 \leq \|\nabla h(y)\| \, d_{\mathcal{T}}^2(x, x'). \tag{36}$$

Proof. Suppose $r \equiv d_{\mathcal{T}}(x, x')$ is less than the injectivity radius on $\mathcal{X}$. Let $\gamma(t)$ be the unit speed shortest path (with respect to tree) connecting x and x' such that $\gamma(0) = x$ and $\gamma(r) = x'$. Let $\widetilde{h}(t) = h(\gamma(t))$. Note that $\widetilde{h}(0) = h(x)$ and $\widetilde{h}(r) = h(x')$. By the mean value theorem, there is some point $0 \leq t' \leq r$ such that

$$\frac{h(x') - h(x)}{d_{\mathcal{T}}(x, x')} = \frac{\widetilde{h}(r) - \widetilde{h}(0)}{r} = \widetilde{h}'(t') = \langle \nabla h(\gamma(t')), \gamma'(t') \rangle.$$

Since γ has unit speed, by the Cauchy–Schwarz inequality, we have

$$|h(x') - h(x)| = |\langle \nabla h(\gamma(t')), \gamma'(t') \rangle| d_{\mathcal{T}}(x, x') \leq \|\nabla h(x')\| \, d_{\mathcal{T}}(x, x').$$

In addition, since $h(\cdot) > 0$ and the unit speed on $d_{\mathcal{T}}$, we have

$$|h^{\frac{1}{2}}(x') - h^{\frac{1}{2}}(x)|^2 \leq \|\nabla h(x')\| \, d_{\mathcal{T}}(x, x') \leq \|\nabla h(x')\| \, d_{\mathcal{T}}^2(x, x').$$

□

Proposition C.6. *There are positive constants $H, I > 0$ such that for $t \in (0, 1]$ and $d_{\mathcal{T}}(x, x') \lesssim t^{\frac{1}{2}}$, we have*

$$|\sqrt{a_t(x, u)} - \sqrt{a_t(x', u)}|^2 \leq H \frac{d_{\mathcal{T}}^2(x, x')}{\sqrt{t} \cdot t^{\frac{n}{2}}} \exp\left(-\frac{I \cdot d_{\mathcal{T}}^2(u, x)}{t}\right). \tag{37}$$

Proof. From Lemma C.6 and Proposition C.5, we have

$$|\sqrt{a_t(x, u)} - \sqrt{a_t(x', u)}|^2 \leq d_{\mathcal{T}}^2(x, x') \frac{E}{\sqrt{t} \cdot t^{\frac{n}{2}}} \exp\left(-\frac{F \cdot d_{\mathcal{T}}^2(u, y)}{t}\right)$$

for some point y lying on the path of $\mathcal{X}$ between x and x'. Because $d_{\mathcal{T}}(x, x') \lesssim t^{\frac{1}{2}}$, the inequality $d_{\mathcal{T}}(x, y) \lesssim t^{\frac{1}{2}}$ also holds. Then, we have

$$d_{\mathcal{T}}^2(u, x) \leq 2\left(d_{\mathcal{T}}^2(u, y) + d_{\mathcal{T}}^2(y, x)\right) \lesssim 2\left(d_{\mathcal{T}}^2(u, y) + t\right)$$

and

$$\begin{aligned}
&|\sqrt{a_t(x, u)} - \sqrt{a_t(x', u)}|^2 \\
\leq & d_{\mathcal{T}}^2(x, x') \frac{E}{\sqrt{t} \cdot t^{\frac{n}{2}}} \exp\left(-\frac{F \cdot d_{\mathcal{T}}^2(u, y)}{t}\right) \\
\lesssim & d_{\mathcal{T}}^2(x, x') \frac{E}{\sqrt{t} \cdot t^{\frac{n}{2}}} \exp\left(-\frac{F \cdot (d_{\mathcal{T}}^2(u, x) - 2t)}{2t}\right) \\
\leq & d_{\mathcal{T}}^2(x, x') \frac{E}{\sqrt{t} \cdot t^{\frac{n}{2}}} \exp\left(-\frac{\frac{F}{2} \cdot d_{\mathcal{T}}^2(u, x)}{t}\right).
\end{aligned}$$

□

Theorem 1. For $0 < \alpha < \frac{1}{2}$ and sufficient K, the multi-scale metric $\hat{T}_\alpha^K$ is equivalent to $d_{\mathcal{T}}^{2\alpha}$.

Proof. By taking $\beta = 2$, Lemma C.4 and Lemma C.5 assure the condition of the upper and lower bounds of the kernel, respectively. Proposition C.6 ensures the condition of the Hölder continuity. Therefore, by applying Proposition C.4 we can obtain the theorem. □

D. Additional Details on the Experimental Study

We present the setups and additional details of the experiments in Section 6. Our code is included in the supplemental material. The experiments are performed on NVIDIA GTX 1080 Ti GPU. A fixed random seed `1234` is used in all the experiments.

D.1. Baselines

The implementation of the competing methods is open-source. The code of tree representation (TR) (Sonthalia & Gilbert, 2020) can be found in the open-source implementation[1]. We use the PyTorch code in (Gu et al., 2018) for Poincaré embedding (PE) (Nickel & Kiela, 2017), the code in (Sala et al., 2018) for the hyperbolic multi-dimensional scaling (hMDS), and PyTorch (PT) code of an SGD-based algorithm, which are all open-source implementations[2]. The code of hyperbolic hierarchical clustering (HHC) (Chami et al., 2020) can be found in the open-source implementation[3]. For the graph embedding task, we also consider a two-dimensional hyperbolic embedding built by Sarkar's combinatorial construction (CC-2) (Sarkar, 2011) and report the hierarchical graph embedding quality in Table 2.

D.2. Datasets

We describe the datasets considered in the experiments in Section 6. They are all publicly available. (i) In the hierarchical graph embedding, the hierarchical datasets considered here are structured as graphs with vertices and edges. Five benchmark datasets in (Sala et al., 2018)[4] are used, including the small balanced tree, the phylogenetic tree, the disease, the CS-PHD, and the Gr-Qc graphs. (ii) In the experiment of scRNA-seq, the datasets are high-dimensional data (samples) measured in an ambient space (gene markers). Two open-source datasets in (Dumitrascu et al., 2021)[5] are considered: Zeisel (Zeisel et al., 2015) and CBMC (Stoeckius et al., 2017). The pre-processing protocol of the scRNA-seq datasets adheres to (Dumitrascu et al., 2021). (iii) In the downstream classification task, four datasets in the UCI Machine Learning repository (Dua & Graff, 2017)[6] are utilized, where the datasets consist of high-dimensional data (instances) collected in an ambient space (attributes). The datasets we used are the Zoo, the Iris, the Glass, and the Image Segmentation datasets, which are used in (Chami et al., 2020) for hierarchical clustering tasks.

D.3. Implementation Details

For the graph embedding task, the diffusion operator is computed by $\mathbf{P} = \exp(-\mathbf{L})$, where $\mathbf{L}$ is the graph Laplacian matrix. This computation is based on the relationship between the diffusion operator and the heat kernel described in Section 2. For high-dimensional data, a distance based on the cosine similarity (`sklearn.metrics.pairwise_distances`) computed in the ambient space is used in Eq. (1), and the diffusion operator is constructed as in Section 2. This distance is also used in the distance-based competing methods, and the corresponding cosine similarity is used in the similarity-based baselines. We compute HDD and the embedding according to Algorithm 1, with the parameter $\alpha = \frac{1}{2}$ and the maximal scale $K \in \{0, 1, \ldots, 19\}$.

Remark. The computation of HDD in Eq. (9) involves the diffusion operator construction, calculating the multi-scale distribution vectors, computing the scaled Hellinger distance between data points, and the summation over the inverse hyperbolic sine functions. It could be computationally heavy when working with large-size datasets (i.e., more than ten thousand data points). The construction of the diffusion kernels is typically the most computationally intensive step. For

[1] https://github.com/rsonthal/TreeRep
[2] https://github.com/HazyResearch/hyperbolics
[3] https://github.com/HazyResearch/HypHC
[4] https://github.com/HazyResearch/hyperbolics/tree/master/data/edges
[5] https://github.com/solevillar/scGeneFit-python/tree/62f88ef0765b3883f592031ca593ec79679a52b4/scGeneFit/data_files
[6] https://archive.ics.uci.edu/ml/datasets.php

large-scale datasets, recent methods in diffusion geometry, such as those presented in (Moon et al., 2019; Tong et al., 2021; Shen & Wu, 2022), have proposed various techniques (downsampling, interpolative approximations, and landmark diffusion, respectively) to significantly reduce the run time and space complexity of diffusion (e.g., $\tilde{O}(mn)$ in (Tong et al., 2021) instead of $\tilde{O}(mn^3)$ and $O(n^{1+2\beta})$ in (Shen & Wu, 2022) instead of $O(n^3)$, where n and m represent the number of samples and features in a data matrix, respectively, and $\beta < 1$ is a hyperparameter related to the size of the landmark set). These techniques can be integrated into HDD, almost as is, enabling the analysis of datasets larger than ten thousand data points using HDD.

E. Additional Experimental Results

E.1. Toy Example

In Fig. 5, we illustrate HDE and HDD on a toy example consisting of a five-level balanced binary tree. In Fig. 5(a), we plot the given tree graph $G = (\mathcal{T}, \mathcal{E}, \mathbf{W})$, where $\mathcal{T} = \{x_i\}_{i=0}^{31}$ is the vertex set organized from the root to the leaves of the tree, $\mathcal{E}$ is the edge set connecting tree nodes, and $\mathbf{W}$ is the edge connectivity matrix. Then, the diffusion operator $\mathbf{P}$ is computed by $\mathbf{P} = \exp(-\mathbf{L})$, where $\mathbf{L}$ is the graph Laplacian of G. An illustration of the multi-scale propagated densities associated with $\mathbf{P}$ and diffusion times in a dyadic grid is shown in Fig. 5(b). We see that the larger the scale k, the more local the support of propagated densities, and the smaller the scale k, the wider the support of the densities. Fig. 5(c) depicts the HDE. Each row represents the multi-scale representation in $\mathcal{H}$, denoted by $\zeta_K(x_i) = \left[(\hat{\mathbf{x}}_i^0)^\top, (\hat{\mathbf{x}}_i^1)^\top, \ldots, (\hat{\mathbf{x}}_i^K)^\top\right]^\top$, of each node. Here as well, we see that as the scale increases (from left to right), the representation becomes more local (concentrating at the diagonal). Fig. 5(d) presents the obtained HDD of each node, where the nodes are colored according to their level in the binary tree. For visualization, we depict the two-dimensional multi-dimensional scaling (MDS) (Cox & Cox, 2008) applied to the nodes using HDD as the input distance. We observe that HDD indeed recovers the tree graph.

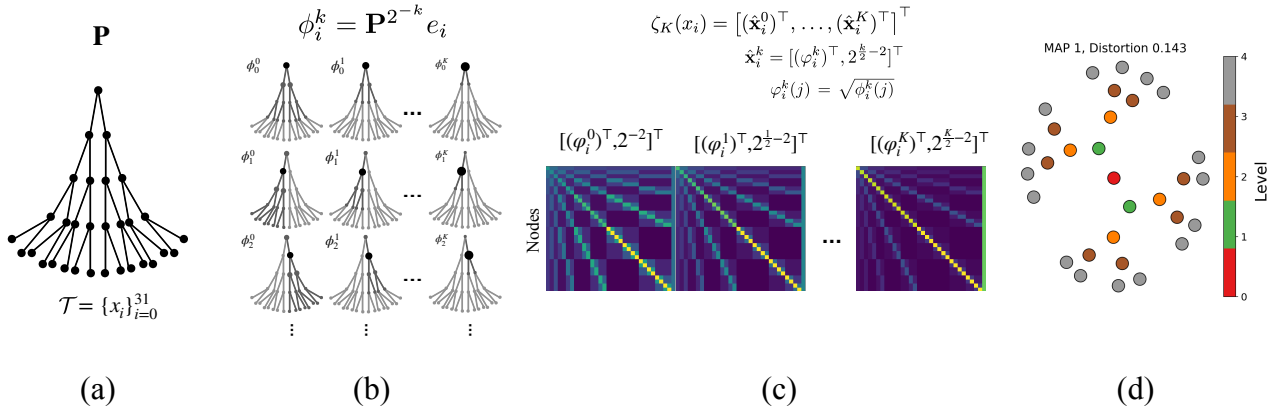

Figure 5. A demonstration of hyperbolic diffusion embedding and distance with an example of a five-level complete binary tree. (a) Given a tree or tree-like graph $G = (\mathcal{T}, \mathcal{E}, \mathbf{W})$, a diffusion operator $\mathbf{P}$ is constructed by the edge connectivity $\mathbf{W}$. (b) The multi-scale propagated densities $\{\phi_i^k\}_{i=0}^{31}$ are computed at each node on $\mathcal{T}$. The rows represent the nodes ordered from the root to the leaves. The columns represent the scale k. The size of the nodes and the color of the edges depict the density value at the nodes and the connectivity between them at scale k, respectively. (c) The HDE of the nodes is plotted in rows from the small scales (left) to the large scales (right). (d) 2D MDS based on HDD. Each point represents a node. The points are colored by the corresponding levels of the binary tree.

E.2. Hierarchical Graph Embedding

We report the obtained MAP and average distortion for the five hierarchical graph datasets in Table 2. The HDD of the balanced tree, the phylogenetic tree, the disease, the CS-PHD, and the Gr-Qc graphs are respectively obtained with $K = 3$, $K = 3$, $K = 3$, $K = 4$, and $K = 10$. We examine the role of maximum scale K in Algorithm 1 in the ablation study in Appendix E.3. We observe that further increasing the maximum scale does not vary the two fidelity measures, indicating the convergence of our proposed method. Table 2 shows that HDD attains a MAP of 1.0 in the small balanced tree and phylogenetic tree, comparable to the combinatorial representation learned from Sarkar's construction (Sarkar, 2011) (CC-2). For tree-like and dense graphs, HDD outperforms the optimization-based approaches PE, PT, and HHC. However, HDD has

larger average distortions than TR, hMDS, PT, and CC-2 for most datasets. We remark that HDD is strictly better in terms of MAP and average distortion than PE and HHC. Arguably, there is a trade-off between the two fidelity measures as noted in (Sala et al., 2018). Our method leans more toward preserving local structure (MAP) at the expense of the global structure (average distortion). In Table. 3, we summarize the execution time for hierarchical representation learning on graphs in this experiment. We observe that HDD is the second or the third fastest algorithm for extracting hierarchical information among the five graph datasets. While HDD is slower than the divide-and-conquer tree representation (TR), the obtained advantage in MAP values shown in Fig. 2 and Table 2 are significant. In addition, we report that HDD is much more efficient than the optimization-based methods: PE, PT, and HHC.

Table 2. MAP and average distortion (D_{avg}) of hierarchical graph embedding.

		HDD	TR	PE-2	PE	hMDS	PT	HHC	CC-2
MAP	Balanced tree	**1.0**	0.942	0.830	0.861	**1.0**	0.964	0.901	**1.0**
	Phylo tree	**1.0**	0.931	0.696	0.724	0.682	0.902	0.884	**1.0**
	Diseases	**0.970**	0.873	0.611	0.912	0.931	0.943	0.831	0.808
	CS-PhD	**0.999**	0.954	0.623	0.781	0.562	0.682	0.774	0.792
	Gr-QC	**0.930**	0.701	0.564	0.763	0.649	0.702	0.685	0.684
D_{avg}	Balanced tree	0.144	0.102	0.446	0.229	0.062	0.131	0.284	**0.010**
	Phylo tree	0.520	0.304	0.841	0.641	0.087	0.207	0.696	**0.009**
	Diseases	0.206	0.187	0.426	0.694	0.123	**0.072**	0.303	0.122
	CS-PhD	0.274	0.194	0.498	0.442	**0.162**	0.243	0.382	0.288
	Gr-QC	0.179	0.202	0.298	0.246	0.542	**0.108**	0.274	0.334

Table 3. Computation time (in seconds) of hierarchical graph embedding.

Dataset	(#Vertices, #Edges)	HDD	TR	PE-2	PE	hMDS	PT	HHC	CC-2
Balanced tree	(40, 39)	$5.89 \cdot 10^{0}$	$\mathbf{4.41} \cdot 10^{-1}$	$1.22 \cdot 10^{2}$	$8.89 \cdot 10^{2}$	$4.92 \cdot 10^{0}$	$9.48 \cdot 10^{2}$	$7.82 \cdot 10^{1}$	$1.63 \cdot 10^{0}$
Phylo tree	(344, 343)	$4.01 \cdot 10^{1}$	$\mathbf{9.83} \cdot 10^{-1}$	$8.74 \cdot 10^{2}$	$1.24 \cdot 10^{3}$	$6.03 \cdot 10^{1}$	$6.33 \cdot 10^{3}$	$1.62 \cdot 10^{2}$	$2.17 \cdot 10^{0}$
Diseases	(516, 1188)	$4.17 \cdot 10^{1}$	$\mathbf{1.02} \cdot 10^{0}$	$1.23 \cdot 10^{3}$	$3.09 \cdot 10^{4}$	$5.21 \cdot 10^{1}$	$1.68 \cdot 10^{4}$	$2.33 \cdot 10^{2}$	$4.42 \cdot 10^{0}$
CS-PhD	(1025, 1043)	$6.44 \cdot 10^{1}$	$\mathbf{1.90} \cdot 10^{0}$	$1.78 \cdot 10^{4}$	$5.62 \cdot 10^{4}$	$9.63 \cdot 10^{1}$	$2.50 \cdot 10^{4}$	$5.43 \cdot 10^{2}$	$7.93 \cdot 10^{0}$
Gr-QC	(4158, 13428)	$9.12 \cdot 10^{1}$	$\mathbf{2.03} \cdot 10^{0}$	$2.89 \cdot 10^{4}$	$3.13 \cdot 10^{5}$	$1.94 \cdot 10^{2}$	$3.43 \cdot 10^{4}$	$1.92 \cdot 10^{3}$	$9.37 \cdot 10^{1}$

E.3. Ablation Study

We conduct an ablation study to investigate the effectiveness of the different components in our method. First, we compare HDD with the ℓ_2 distance in the product manifold $\mathcal{H}$, given by

$$d_{\mathcal{H}}^{\ell_2}\left(\zeta_K(x_i), \zeta_K(x_{i'})\right) = \sum_{k=0}^{K} \left(2\sinh^{-1}\left(2^{-k\alpha+1}\left\|\varphi_i^k - \varphi_{i'}^k\right\|_2\right)\right)^2, \tag{38}$$

where $K \in \mathbb{Z}_0^+$ is the maximum scale defined in the same way as in HDD. Note that $d_{\mathcal{H}}^{\ell_2}$ is equipped with a Riemannian structure (Ficken, 1939). In addition, we test single scales in the factor manifold $\mathbb{H}^{n+1}$ in the product manifold $\mathcal{H}$, given by

$$d_{\mathbb{H}^{n+1}}(\hat{\mathbf{x}}_i^k, \hat{\mathbf{x}}_{i'}^k) = 2\sinh^{-1}\left(2^{-k\alpha+1}\left\|\varphi_i^k - \varphi_{i'}^k\right\|_2\right), \tag{39}$$

where $k \in \mathbb{Z}_0^+$ represents the scale.

The results, comparing HDD, $d_{\mathcal{H}}^{\ell_2}$, and the single scale embedding are presented in Fig. 6 for the graph embedding experiment presented in Section 6.1. The five plots depict the distortion-MAP graph for the five datasets. In each plot, the blue circle, green plus, and red star represent the result of HDD, $d_{\mathcal{H}}^{\ell_2}$, and the single scale embedding, respectively. The color of the points represents the parameter K (resp. k) for HDD and $d_{\mathcal{H}}^{\ell_2}$ (resp. single scale embedding). Note that when $k = K = 0$, HDD and the single scale embedding coincide (i.e., $d_{\mathcal{H}}^{\ell_1}\left(\zeta_0(x_i), \zeta_0(x_{i'})\right) = d_{\mathbb{H}^{n+1}}(\hat{\mathbf{x}}_i^0, \hat{\mathbf{x}}_{i'}^0)$). We observe that HDD outperforms the other two alternatives, indicating that, indeed, the use of the ℓ_1 norm and the multiple scales in Eq. (9) has a critical

contribution to the extraction of the hierarchical structure, as guaranteed in Theorem 1. In addition, based on the results of HDD and $d_{\mathcal{H}}^{\ell_2}$, we find that the larger K is, the better the embedding quality is. Conversely, the role of k plays an opposite effect in the single embedding, as conveyed in Proposition 1. Last, we see that the results of HDD in terms of MAP and average distortion converge for sufficiently large K, providing empirical support to the approximation in Eq. (21).

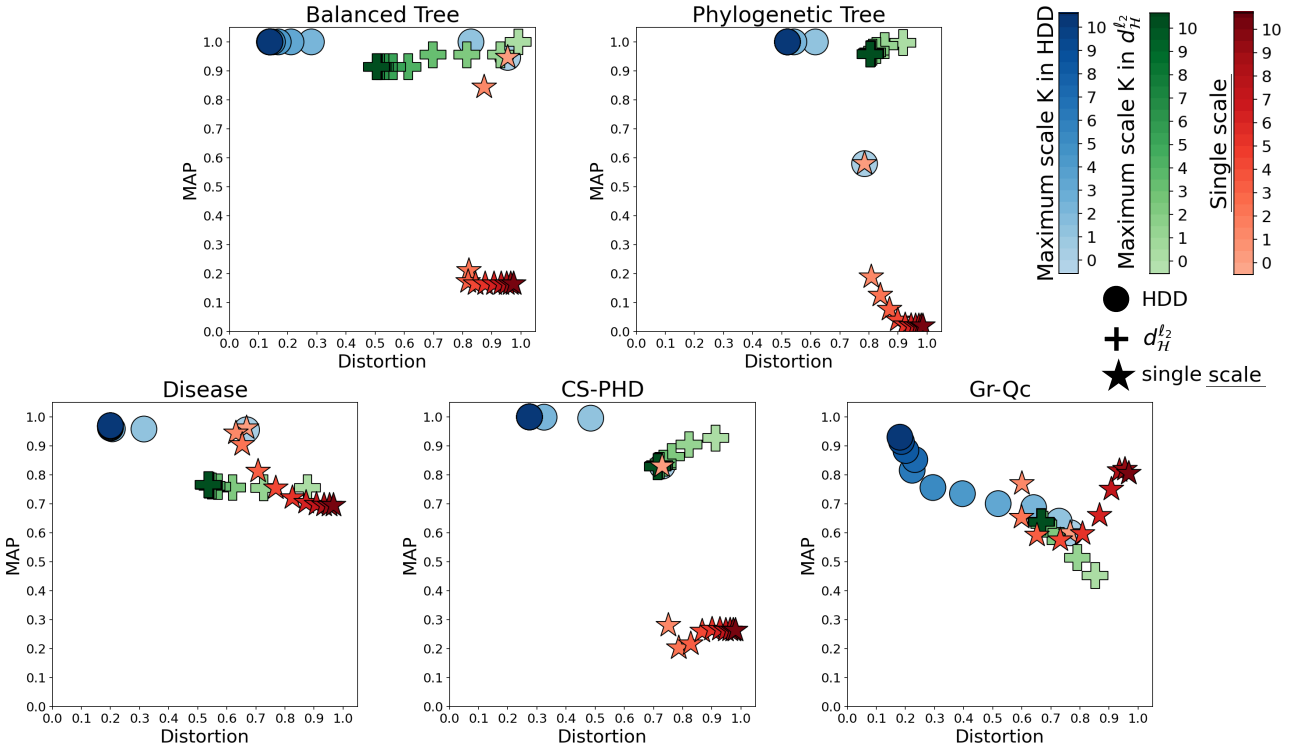

Figure 6. Distortion-MAP plots of the five datasets for graph embedding. In each plot, the blue circle, green plus, and red star represent the result of HDD, $d_{\mathcal{H}}^{\ell_2}$, and the single scale embedding, respectively. The color of the points represents the parameter K (resp. k) for HDD and $d_{\mathcal{H}}^{\ell_2}$ (resp. single scale embedding).

In addition, we conducted experiments that compare the performance of the proposed HDD with a variant in which the hyperbolic distance is replaced by the following Euclidean distance

$$d_{\text{Euc}}(i, i') = \|\zeta_K(x_i) - \zeta_K(x_{i'})\|_2 \,. \tag{40}$$

Our results are presented in Table 4, where we can see that using the Euclidean distance does not capture the hierarchical structure. This empirical evidence demonstrates the importance of the hyperbolic distance in our method, showing that the proposed construction of HDD is essential to the recovery of the hierarchy.

Table 4. MAP and average distortion (D_{avg}) of hierarchical graph embedding using the Euclidean distance.

		Balanced tree	Phylo tree	Diseases	CS-PhD	Gr-QC
MAP	HDD	1.0	1.0	0.970	0.999	0.930
	Euc	0.219	0.154	0.132	0.116	0.228
D_{avg}	HDD	0.144	0.520	0.206	0.274	0.179
	Euc	0.694	0.712	0.736	0.688	0.781

E.4. Single-Cell Gene Expression Data

The obtained MAP, average distortion, and classification accuracy of the scRNA-seq datasets are reported in Table 5. In the gene expression data, the maximum scales used in Algorithm 1 for the Zeisel and the CBMC datasets are set to $K = 9$ and $K = 13$, respectively. We note that they are slightly larger than the scales used in the graph embedding experiment due to the larger size and dimensionality of the data. Observing the table, we see that HDD achieves the best MAP and the second-best average distortion. In terms of classification accuracy, HDD outperforms all the competing methods in both scRNA-seq datasets. We report the run time of HDD and the competing baselines in Table 6. Note that the optimization-based methods, PE, PT, and HHC, require a much longer time to find the hierarchical representation, similar to the hierarchical graph embedding task in Table 3. Yet, the additional computational time does not lead to improved embedding quality and downstream classification accuracy. Our HDD obtains a slightly larger distortion than TR, and it is slower than TR. Yet, its advantage in terms of MAP and classification accuracy is significant, as illustrated in Fig. 3, Fig. 4, and Table 5.

Table 5. MAP, average distortion (D_{avg}), and classification accuracy of scRNA-seq datasets.

		HDD	TR	PE-2	PE	hMDS	PT	HHC
Zeisel	MAP	**0.996**	0.803	0.779	0.788	0.710	0.542	0.853
	D_{avg}	0.169	**0.121**	0.278	0.223	0.213	0.581	0.482
	ACC.	**0.862**$\pm$**0.014**	0.664$\pm$0.039	0.712$\pm$0.018	0.743$\pm$0.018	0.802$\pm$0.041	0.597$\pm$0.098	0.811$\pm$0.039
CBMC	MAP	**0.979**	0.713	0.749	0.817	0.789	0.760	0.806
	D_{avg}	0.297	**0.254**	0.522	0.489	0.364	0.473	0.323
	ACC.	**0.832**$\pm$**0.023**	0.741$\pm$0.037	0.739$\pm$0.061	0.752$\pm$0.019	0.733$\pm$0.039	0.648$\pm$0.027	0.788$\pm$0.029

Table 6. Execution time (in seconds) of scRNA-seq datasets.

Dataset (#Points, #Classes)		HDD	TR	PE-2	PE	hMDS	PT	HHC
Zeisel	(3005, 7)	$1.13 \cdot 10^2$	$\mathbf{1.09} \cdot 10^0$	$2.13 \cdot 10^4$	$2.21 \cdot 10^4$	$1.07 \cdot 10^2$	$1.07 \cdot 10^4$	$1.69 \cdot 10^3$
CBMC	(8617, 13)	$5.91 \cdot 10^2$	$\mathbf{2.48} \cdot 10^0$	$4.26 \cdot 10^4$	$4.66 \cdot 10^4$	$4.13 \cdot 10^2$	$3.73 \cdot 10^4$	$4.54 \cdot 10^3$