UNIVERSITY OF CALIFORNIA SAN DIEGO

Hyperbolic geometry in biological systems

A dissertation submitted in partial satisfaction of the
requirements for the degree Doctor of Philosophy

in

Biology with a Specialization in Quantitative Biology

by

Yuansheng Zhou

Committee in charge:

Professor Tatyana Sharpee, Chair
Professor Terry Sejnowski, Co-Chair
Professor Henry Abarbanel
Professor Takaki Komiyama
Professor Pamela Reinagel

2021

The Dissertation of Yuansheng Zhou is approved, and it is acceptable in quality and form for publication on microfilm and electronically.

University of California San Diego

2021

DEDICATION

This dissertation is dedicated to my thesis advisor Dr. Tatyana Sharpee who inspired and mentored the projects.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to pay my special regards to my thesis advisor Dr. Tatyana Sharpee who was wise and patient in mentoring my research, caring for my needs, and always supportive for my decisions.

I wish to express my deepest gratitudes to Pastor Chris and my Christian friends in church – Simon, Aaron, Jeff, Yaohui, Sufren, Christy and Jenny, who gave me spiritual support and company during these years, especially my girl friend Xuewei and my parents who always love, care and trust me.

I would like to thank all the people who played significant roles in my journey of research, including my committee members who kept me on track and gave insightful comments, and my lab mates who were precious supporters in my research.

I wish to show my gratitude to the organizations that support my research: Biological Sciences Program in UCSD, the Salk Institute, and NSF which funds my research.

Chapter 1, in full, is a reprint of the material as it appears in Hyperbolic geometry of the olfactory space. *Science advances*, 2018; Yuansheng Zhou, Brian H Smith, Tatyana O Sharpee. The dissertation author was the primary investigator and author of this paper.

Chapter 2 , in full, is a reprint of the material as it appears in Hyperbolic geometry of gene expression. *Iscience*, 2021; Yuansheng Zhou, Tatyana O Sharpee. The dissertation author was the primary investigator and author of this paper.

Chapter 3, in full, is currently being prepared for submission for publication of the material. Yuansheng Zhou, Tatyana O Sharpee. The dissertation author was the primary investigator and author of this material.

VITA

2015	Bachelor of Science, Peking University, Beijing, China
2021	Doctor of Philosophy, University of California San Diego, San Diego, USA

PUBLICATIONS

Yuansheng Zhou and Tatyana O Sharpee. Hyperbolic geometry of gene expression. *Iscience*, 24(3):102225, 2021.

Debha N Amatya, Sara B Linker, Ana PD Mendes, Renata Santos, Galina Erikson, Maxim N Shokhirev, Yuansheng Zhou, Tatyana Sharpee, Fred H Gage, Maria C Marchetto, Yeni Kim. Dynamical Electrical Complexity Is Reduced during Neuronal Differentiation in Autism Spectrum Disorder. *Stem cell reports*, 2019; 13 (3), 474-484

Yuansheng Zhou, Brian H Smith, Tatyana O Sharpee. Hyperbolic geometry of the olfactory space. *Science advances*, 2018; 4(8):eaaq1458.

Yuansheng Zhou, Tatyana Sharpee. Using global t-SNE to preserve inter-cluster data structure. bioRxiv. 2018; p. 331611.

FIELDS OF STUDY

Major Field: Biology with a Specialization in Quantitative Biology

ABSTRACT OF THE DISSERTATION

Hyperbolic geometry in biological systems

by

Yuansheng Zhou

Doctor of Philosophy in Biology with a Specialization in Quantitative Biology

University of California San Diego, 2021

Professor Tatyana Sharpee, Chair
Professor Terry Sejnowski, Co-Chair

Many biological systems have intrinsic hierarchical structure, which can be best described by hyperbolic geometry. Hyperbolic geometry was well developed in mathematics, and has recently been introduced in machine learning for data representation and visualization. However, the relevance of hyperbolic geometry to biological data was rarely studied. In this dissertation, I develop methods to systematically detect the existence of hyperbolic geometry in various biological systems, and show the advantages of hyperbolic representation of biological data. The geometry detection of these data are performed using either the sensitive and accurate Betti curve method or the robust and fast multi-dimensional scaling method. It turns out that many biological systems, including olfactory data from different species and gene expression profiles from different sources, all have hyperbolic structure. The visualization of the data can be achieved by either hyperbolic multi-dimensional scaling or hyperbolic t-SNE, which are both proposed in this dissertation. The former method correctly reveals the global organization of data points and identifies the phenotype related axes in the space, e.g. the pleasantness axis in olfactory space. The latter method aims to perform local clustering as well as preserving inter-cluster structure, giving an accurate representation of cell types in an informative way. In addition to the improved data visualization, hyperbolic representations of data inspire new discoveries that cannot be obtained by traditional Euclidean-based methods, for example, different cell types in brain are characterized by distinct spatial localization patterns in the hyperbolic disk, regardless of brain regions.

Introduction

Modern datasets characterize natural objects with respect to many variables and assign distances between objects based on a Euclidean metric. For example, the plants can be described by a number of traits, the cells are classified by the measurements of the transcriptome, and the odor perception can be characterized by a set of human descriptors, etc. The high-dimensional measurements are primarily evaluated by a Euclidean metric since it's the best prior assumption for the high-dimensional data. However, recent results suggest that for data produced by underlying hierarchical tree-like networks a hyperbolic metric might be more appropriate than a Euclidean one. So what is the real geometry that a high-dimensional data represent, and how to properly visualize it? These are the two questions this dissertation aims to address. Firstly, I develop two approaches to detecting the hidden geometry of different biological systems, which are all described by high-dimensional measurements. One of the approaches is Betti curve method which is fit for small datasets and can precisely determine the parameters of geometry, the other one is non-metric multi-dimensional scaling (MDS) method which is robust and can be applied to data of any size. Secondly, I propose two visualization algorithms that can embed data into two dimensional hyperbolic disks. These visualization algorithms include hyperbolic MDS (HMDS) method which aims to reveal global organization of data in the space, and hyperbolic t-SNE (h-SNE) method which performs local clustering as well as preserving the global inter-cluster orientations. I show that the geometry-preserving embeddings of hyperbolic data significantly outperform traditional Euclidean-based methods in accuracy, and reveal more biological insights.

In Chapter 1, I apply Betti curve method to show that the geometry of olfactory space is hyperbolic and determine the geometric parameters for both odor stimuli space and odor perception space. I use HMDS method to visualize the data in the identified geometry and find three phenotype-related axes in the odor stimuli space. In Chapter 2, I develop a methodology for geometry detection of large gene expression data and find the gene expression space from five different sources all have hyperbolic structure. I propose hyperbolic t-SNE method to visualize the data in 2D hyperbolic disk, and show that it significantly outperforms other methods quantitatively and qualitatively. In Chapter 3, I extend the application of HMDS and h-SNE methods to three other datasets and show several preliminary results in which these new methods give novel biological insights and inspire interesting future directions.

Chapter 1

Hyperbolic geometry of olfactory space

1.1 Abstract

In the natural environment, the sense of smell, or olfaction, serves to detect toxins and judge nutritional content by taking advantage of the associations between compounds as they are created in biochemical reactions. This suggests that the nervous system can classify odors based on statistics of their co-occurrence within natural mixtures rather than from the chemical structures of the ligands themselves. We show that this statistical perspective makes it possible to map odors to points in a hyperbolic space. Hyperbolic coordinates have a long but often under-appreciated history of relevance to biology. For example, these coordinates approximate the distance between species computed along dendrograms and, more generally, between points within hierarchical tree–like networks. We find that both natural odors and human perceptual descriptions of smells can be described using a three-dimensional hyperbolic space. This match in geometries can avoid distortions that would otherwise arise when mapping odors to perception.

1.2 Introduction

The reason that the sense of smell can be used to avoid poisons or estimate a food's nutrition content is because biochemical reactions create many by-products. Thus, the emission of certain sets of volatile compounds will accompany the production of a specific poison by a plant or bacteria. An animal can therefore judge the presence of poisons in the food by how the food smells. Other specific examples include the use of smell by bees when judging whether a flower has more pollen or nectar [1, 2]. Fruit flies select places to lay eggs based on odors[3]. These examples suggest that, from a practical perspective, it would be useful for the nervous system to classify odors based on statistics of their co-occurrence. For example, if odor components that are strongly correlated are represented nearby within the nervous system [4], then detection of one component could be quickly used as an indicator for the likely existence of another component that is strongly correlated with it. With this perspective in mind, we set out to study the structure of the olfactory space based on odor co-occurrence. Before we describe the results, we review the reasons for why one might expect to find hyperbolic coordinates to be relevant for olfaction and biological systems in general. Biological data are often represented using dendrograms or hierarchical tree structures (Fig. 1A). These data can be equivalently represented using Venn diagrams, where larger circles correspond to broader classifications (Fig. 1B)[5]. For example, before Darwin, these Venn diagrams were used to classify species based on their properties [6]. Darwin used the mapping from Venn diagrams to trees [6] to infer the likely tree for speciation based on available descriptions of species properties (Venn diagrams). There is a deep mathematical reason underlying the equivalence between these two representations, and it involves hyperbolic spaces. Specifically, starting with the Venn diagram (Fig. 1B), one can assign points to a three- dimensional (3D) space whose horizontal x and y coordinates equal to center coordinates of the Venn circles, whereas the vertical coordinate equals to the circle radius [7]. In this manner, larger circles get assigned to higher heights, which would then correspond to positions closer to the tip of the tree (Fig. 1C). Sometimes, the presence of partially overlapping circles leads to a structure that is not precisely a tree because it contains loop. Nevertheless, the resulting 3D space has a hyperbolic metric [8] and can be described by the Poincare half-space model for the hyperbolic space. The fact that the metric is non-Euclidean can be ob- served from the fact that the shortest distance between two points goes up in the z-direction (along the tree) before descending back to the tar- get node. In Fig. 1D, we show an example shortest path between two points in a 2D half-space model (red dashed line) and its discrete approximation (red solid line). To foreshadow the results on olfactory odor classification, we note that 3D hyperbolic space is the lowest dimensional space where the descriptor sets (Venn diagrams) are not 1D, as in Fig. 1D, but are 2D circles as in Fig. 1 (B and C). At least two axes have been described for the human odorant perception (the "pleasantness-to- unpleasantness" axis and the "chemical-to-natural" axis) [9] [10]. Together, these mathematical and biological observations point to the relevance of 3D hyperbolic geometry for odor perception.

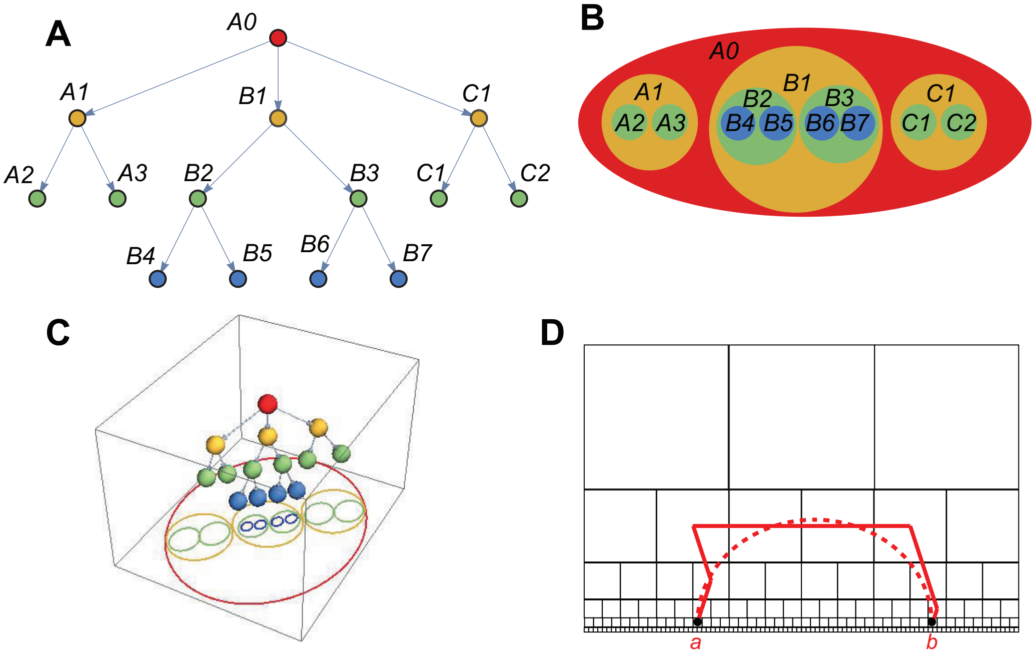

Figure 1. Hyperbolic spaces approximate hierarchical networks. (A) Example hierarchical description of data and (B) its equivalent representation using Venn diagrams. (C) Venn diagrams can be mapped onto points in a 3D space, forming approximately a tree. The metric in the resulting 3D space is hyperbolic [7]. The hyperbolic aspects of the metric are illustrated by the fact that the shortest path be- tween nodes in the tree goes upward and then descends back to the target node. (D) Discrete approximation to a half-space model of the hyperbolic space in 2D. Red solid and dashed lines show the discrete and continuous shortest paths between point *a* and *b* within the half-space model of the hyperbolic space [8].

1.3 Results

1.3.1 Geometry of odor stimuli space

To analyze which space best describes the statistics of co-occurrence within natural odor mixtures, we used a recently developed a statistical method [11] that can identify the presence of a geometric structure in data based on observed correlations between data components. This method is unaffected by linear or nonlinear monotonic transformations of inputs and therefore can be used to determine the overall geometry of the data without worrying at first about the precise scaling of the axes. The analysis starts by taking a set of measurements of concentrations of individual monomolecular odors, as they occur in the natural environment. Our analyses will be based on four data sets of odors measured from samples of strawberries [12], tomatoes [13], blueberries [14], and mouse urine [15]. To give an overview of the data, 69 monomolecular odors were measured across 50 different mouse urine samples, 66 monomolecular odors across 79 tomato samples, 45 monomolecular odors across 101 blueberry samples, and 78 monomolecular odors across 54 samples of strawberries. The first step of the analysis is to compute correlations between the concentrations of monomolecular odor across samples (Fig. 2A). The correlation coefficient between two odors x, y was defined as

$$corr(x,y) = \frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{1.1}$$

where x_i and y_i represent concentrations of odor x and odor y in i_{th} sample, $\bar{x}$ and $\bar{y}$ represent the mean values of concentrations across all samples, and n is the number of samples. Correlation coefficients were computed separately for each data set. The absolute values of the correlation matrix from each data set are passed through a step function with different thresholds: The correlation values above the threshold are set to 1, and the rest of values are set to 0 (Fig. 2B). The transformed matrix can now be visualized as a topological graph where all unit values represent links between the corresponding odors (Fig. 2C). The graph can be characterized by the number of holes (cycles) in one, two, or higher dimensions. For high thresholds, the number of cycles will be low because most units are not connected. Similarly, at low thresholds, the number of cycles is also low because units form fully connected networks. Plotting the number of cycles as a function of density of edges, or equivalently the number of connected nodes, yields the so-called Betti curves. It turns out that the shapes of these Betti curves are quite sensitive to the statistics of correlations. This sensitivity makes it possible to infer the geometry of the space that can produce these correlations if we sample points from this space and assume that stronger correlations (before thresholding) imply closer distances. [11]

Applying this statistical approach to each of the four data sets separately, we found the data in each case to be consistent with being drawn from a neighborhood of a sphere positioned within a 3D hyperbolic space together with a small amount of multiplicative noise added to the distances (Fig. 2D). The fact that hyperbolic space approximates hierarchical tree–like networks motivated this choice of the model [7], with odors reflecting leaves of the network—the neighborhood of the surface (Fig. 1). Quantitatively, one can compare Betti curves derived from a model geometrical space and from a data set by computing the integral of the curve [11], the quantity referred to as the integrated Betti value. To find the best-fitting geometry, we optimized parameters of the model such that the noise magnitude and the range of radii within the space from which the sample points were drawn provided the best match to the first integrated Betti value. Then, we examined how these optimized parameters could account for the second and third integrated Betti values. For all four data sets, we found the measurements to be consistent with sampling from a 3D hyperbolic space ($P > 0.25, P > 0.21, P > 0.45$, and $P > 0.19$ for blueberry, tomato, mouse, strawberry data sets, respectively; in each case, the value stated is the minimal value across the Betti curves in 2D and 3D; see also table S1). The first three Betti curves were also sufficient to show that Euclidean spaces could not account for the data, even when dimensionality and other parameters were optimized (Fig. 2E, $P < 0.03$ for blueberry and $P < 0.003$ for other three data sets). As a control, we verified that shuffling odor concentrations between samples, which destroys correlations between odors, produced Betti curves that can

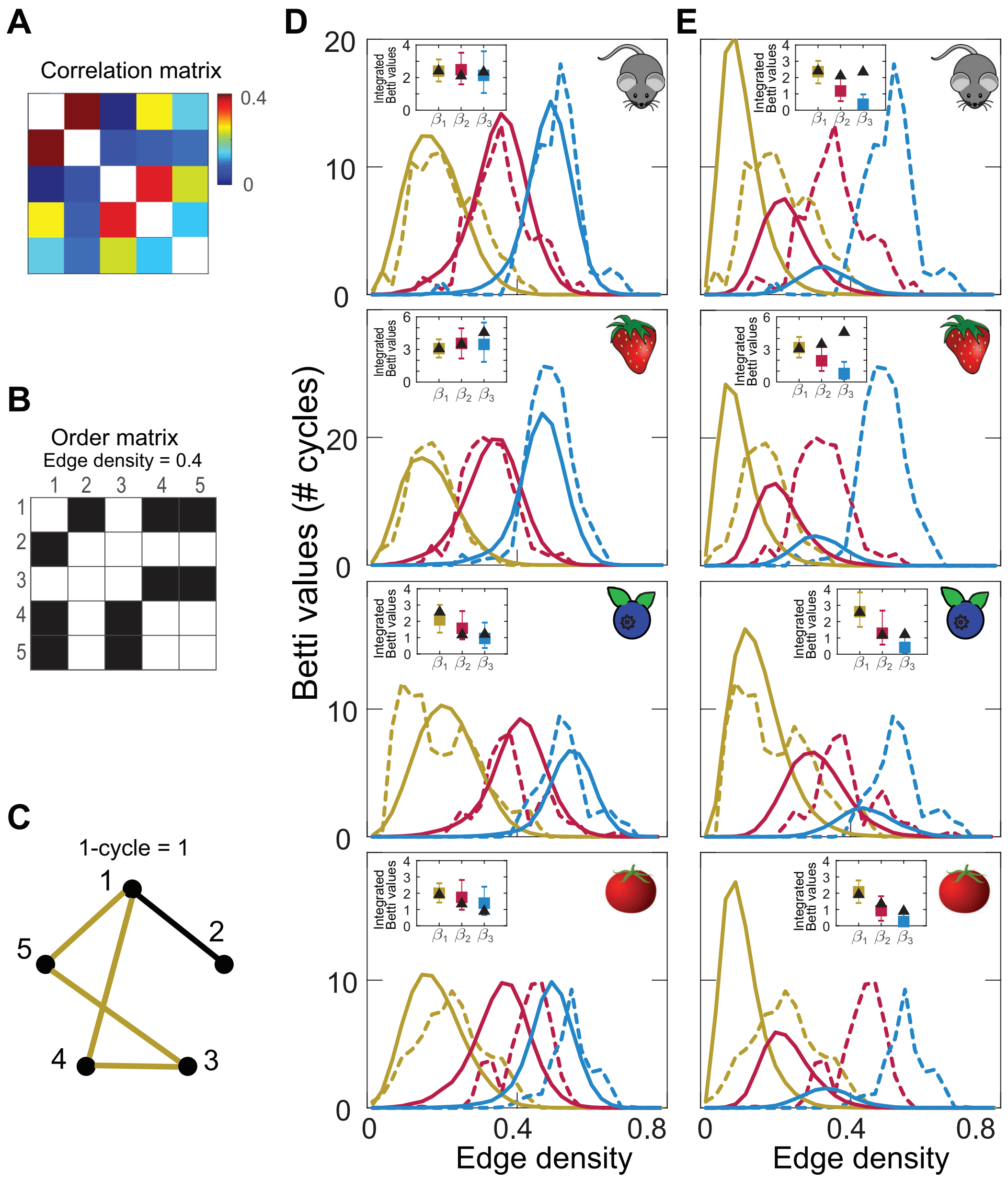

Figure 2. Topological organization of the natural odor space. (A to C) Illustration of the topological algorithm for identifying spaces consistent with correlation statistics. (A) Example correlation matrix for five odors in strawberry data set. (B) Correlation matrix after applying a threshold of 0.25. (C) A nonzero value represents an edge connecting the two elements. The resulting complex has one 1D cycle and edge density of 0.5. (D and E) Betti curves with the number of cycles in one (yellow), two (red), and three (blue) dimensions plotted as a function of edge densities. Data from Betti curves (dashed) are compared with predictions using model geometry (solid lines) of 3D hyperbolic space in (D) or Euclidean space (E). Insets show comparisons between integrated Betti values from data (black triangles) compared with models. The error bars show 95% confidence intervals (from 2.5% to 97.5%) from 300 models with the same number of odors as the data, and the colored squares show the medium values of the models.

be fully ex- plained by random matrices ($P = 0.4, P = 0.7$, and $P = 0.9$ for integrated Betti values one through three, respectively; cf. Fig. S1). These matrices would not be consistent with the hyperbolic space plus the small amount noise that fits the real data ($P < 0.01$). As additional controls, we verified that (i) evaluating differences between Betti curves using L1 distances instead of the integrated Betti values (tables S2 and S3) or (ii) applying logarithm to concentration values before computing their correlations led to the same conclusions (Fig. S2 and table S4). In particular, hyperbolic 3D is consistent with measurements for all three Betti curves, whereas the best-fitting Euclidean model can be ruled out according to these measures. The corresponding P values are provided in tables S1 to S4. Note that hyperbolic spaces of dimensions higher than three cannot be ruled out (Fig. S3). However, the 3D hyperbolic space remains the best-fitting model across the four data sets. This is true whether one uses either the integrated Betti value or the L1 distances between model and experimental Betti curves (Fig. S3).

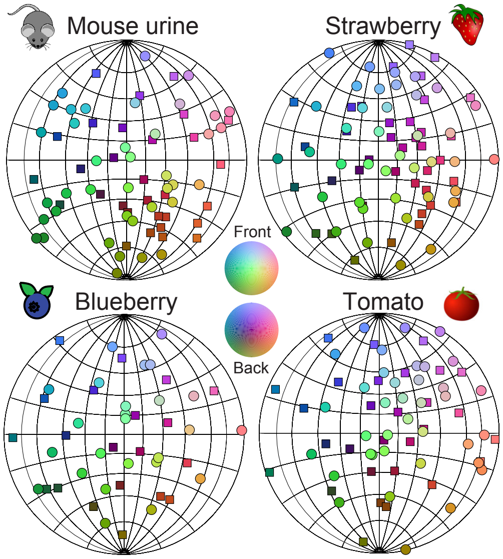

Figure 3. Visualization of the natural olfactory space using non-metric MDS. Be- cause the variation in radius is small, data points are shown on the surface of a sphere with circles/rectangles for points falling on the near/far side of the sphere. The RGB color scales were proportional to the XYZ coordinates of points.

1.3.2 Visualization in hyperbolic space

To visualize how the points consistent with odorant correlation statistics might be distributed within the hyperbolic space, we used non- metric multidimensional scaling (MDS) [16]. The non-metric MDS algorithm embeds a set of points into the N-dimensional space while attempting to preserve the rank ordering of distances as best as possible ([16]). Traditionally, MDS is applied to the Euclidean space, but we modified it (see Materials and Methods) to work with hyperbolic distances [7]. After testing the algorithm on synthetic data (Fig. S4), we applied the modified algorithm to the four data sets. In Fig. 3, we show results for the four data sets. Because the points are located near a surface of a sphere (the range of radii $R_{min} = 0.9R_{max}$), we present the points on a sphere using the two angles of latitude and longitude. The results show approximately uniform sampling in all four data sets. Notably, the points do not cluster based on functional chemical properties of the individual components (Fig. S5). One can understand the absence of clustering from the fact that monomolecular odors with different func- tional properties are produced together in biochemical pathways.

What are the axes of this olfactory space? Or, in other words, do odors associated with certain parts of this space have different perceptual or physicochemical properties? Previous studies found axes that correlated with perceptual odor pleasantness [17] and physicochemical properties such as molecular boiling point and acidity ([9][17][18]). We checked for associations with all of these properties, and supporting previous findings ([9][10][17][19]) found that points corresponding to pleasant and unpleasant odors occupied different parts of the space. We note that this analysis of pleasantness rankings was based only on odor components from tomato and strawberry data sets, for which these rankings were available. Thus, the pleasant-unpleasant odor axis can be identified even solely using fruit odor components. The direction most associated with a change in pleasantness value is marked by the red line in Fig. 4 (A and C). For odor mixtures produced by individual fruit samples, we use the "overall liking" rating assigned by humans to fruit samples as a measure of pleasantness (Fig. 4A). To test how well the identified

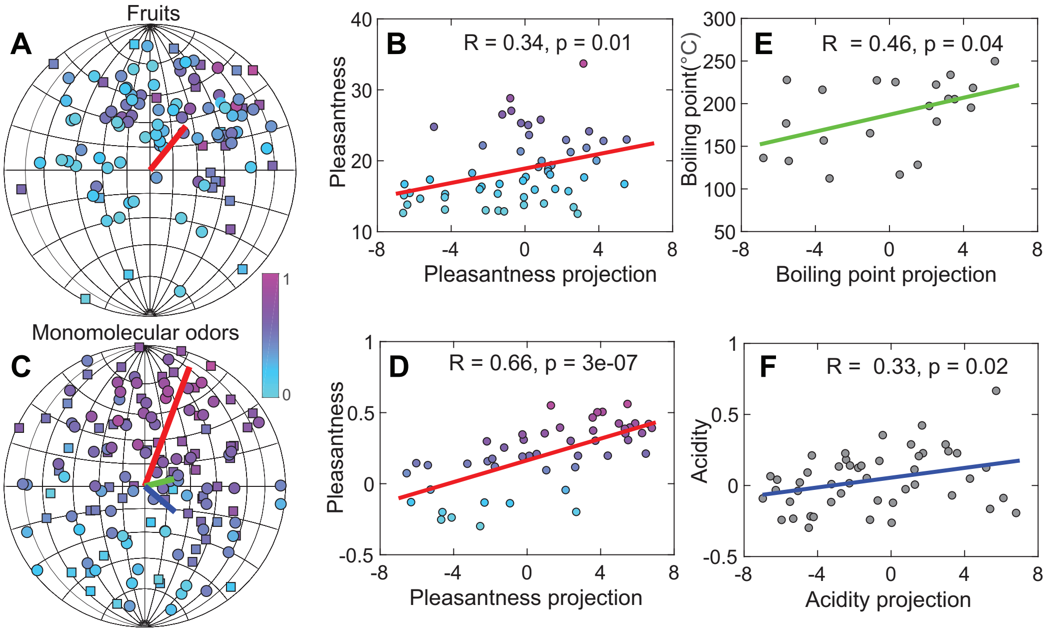

Figure 4. Axes associated with pleasantness and odor physiochemical properties. (A) We represented 108 fruit samples (54 tomatoes and 54 strawberries) using normalized linear combinations of odor coordinates in the space, with the weights proportional to odor concentrations in the samples. The color indicates human rating for the overall liking; circles/squares represent points from the front/back sides of the sphere. The red line shows the direction most associated with the pleasantness ratings. (B) The correlation is significant, with correlation value $R = 0.34$ and $P = 0.01$. (C) Visualization of 144 individual monomolecular odors. The red, green, and blue lines showed the directions of pleasantness, boiling point, and acidity, respectively. Two-thirds of monomolecular odors were used to determine the directions associated with perceptual or physicochemical properties of odors, and the rest one-third were projected onto the directions as validation sets to evaluate the correlations. In the 144 monomolecular odors, only 62 of them have available boiling points and were used to find the boiling points direction. (D) Correlation between odor pleasantness with projection onto the pleasantness axes. (E) Correlation between molecular boiling point, a measure of odor volatility, with projection on the associated axes. (F) Correlation between acidity value and the associated axes.

pleasantness axes can predict measured pleasantness rankings for novel samples, we regenerated this axis using only strawberry samples and use it to predict pleasantness ranking for tomato samples. The correlation was significant, with correlation coefficient $R = 0.34$ and $P = 0.01$ (Fig. 4B). The pleasant- ness values could also be assigned to individual odor components based on the correlation between the odor concentration in a mixture and mixture pleasantness computed over samples (Fig. 4C). This measure of pleasantness produced an even stronger correlation between pleasantness and odor coordinates within the space,$R = 0.66$ and $P = 3 \times 10^{-7}$ (Fig. 4D). We also computed these correlation values using different odor components from those used to generate the pleasant-unpleasant axis for odor components in Fig. 4C. Specifically, we used two-thirds of randomly selected odor components to generate this axis; we computed the correlation value using the remaining components. The pleasantness axis for odor components had a similar orientation to the one for mixtures (that is, individual fruit samples).

In addition to the pleasantness axis, we could also find axes that were strongly associated with two other properties: molecular boiling point—which is probably a reflection of volatility—and acidity, both of which showed significant correlations ($P < 0.04$; Fig. 4, E and F). We assigned acidity for individual odors as the correlation coefficient be- tween its concentration and fruit sample acidity measurement. (We computed all of these correlation coefficients using a different subset of odors from the one used to estimate the corresponding axes.) Be- cause the space is essentially 2D, the three axes of odor pleasantness, acidity, and molecular boiling point are not independent. In other words, knowing the coordinates along the molecular boiling point and acidity axes, one can predict the position along the pleasantness axis. That is, the identified mapping to a sphere in a hyperbolic space makes it possible to predict, with correlation $R = 0.34$ (Fig. 4B) for natural mixtures and with $R = 0.66$ (Fig. 4D) for monomolecular odors, how perceptually pleasant these odors are based on their projections on the acidity and volatility axes.

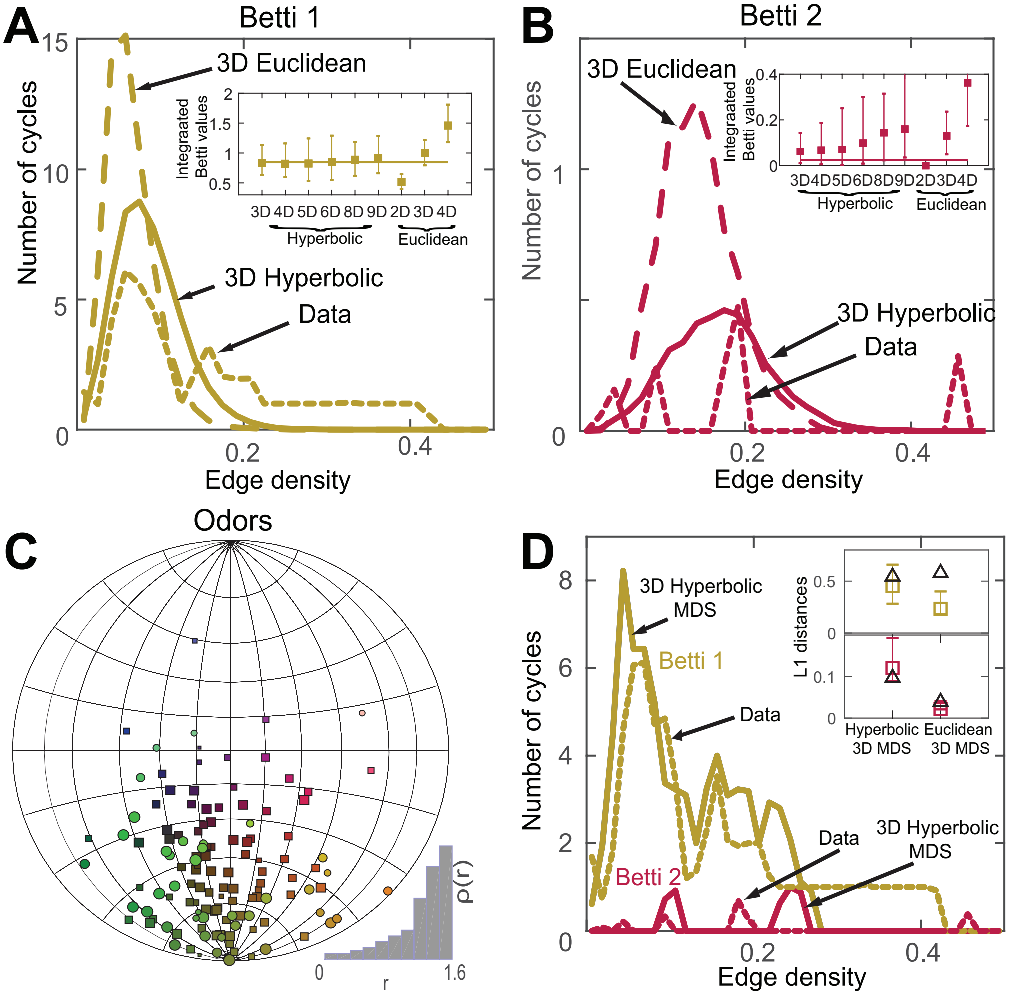

Figure 5. Hyperbolic organization of the human olfactory perception. The (A) first and (B) second Betti curves of the perceptual data set (dotted line) (20) compared to Betti curves of the 3D hyperbolic space (solid line) and 3D Euclidean space (dashed line). Euclidean and hyperbolic spaces of other dimensions provided a worse fit. Insets compare integrated Betti values from data (horizontal lines) and 300 repeated models in different dimensions with Euclidean or hyperbolic metrics. The error bars show 95% confidence intervals; the number of repeated computations of model curves was 300. (C) Visualization of odors in human olfactory perception space using non-metric MDS in a 3D hyperbolic space. The sizes of points are proportional to their radii. The radius distribution is shown in bottom right inset. (D) The multimodal aspects of Betti curves derived from data (dotted lines) can be accounted for by the nonuniform distribution of points within the 3D hyperbolic space. Sampling points from (C) produces multimodal first (yellow) and second (red) Betti curves (solid lines). Inset shows comparison of L1 distances between Betti curves derived from data and those derived from 100 different MDS fits. Black open triangles represent the distance between data and model mean, and colored bar plots show the range of values, where data curves are substituted by different MDS fits.

1.3.3 Geometry of odor perception space

The observation that the odor mixtures can be mapped onto a continuous metric space is consistent with a previous vector-based model of human olfactory perception ([19]). This model posits that the perception of odor mixtures is based on a combination of the mixture components or, in other words, that there is an underlying set of coordinates that can represent olfactory mixtures. Previous analysis ([9]) of the Dravnieks database ([20]) containing human perceptual descriptions of > 120 monomolecular odors showed that the perceptual space is likely to be curved. Qualitatively, the points were found to form a "potato-chip" surface ([9]). This can be a signature of the hyperbolic space; potato-chip or saddle-like surface have a negative curvature and serve as an everyday example of hyperboloid surfaces. To quantitatively test whether the perceptual space is described by hyperbolic geometry, we applied the Betti curve method to the Dravnieks database [20]. First, we found that Euclidean spaces were not consistent with measured Betti curves ($P < 0.003$;Fig. 5A and table S5). The first Betti curve could not be matched to the data in terms of its area for any dimensionality of the Euclidean space (Fig. 5A, inset). Second, we found that the full hyperbolic space of varying dimensions could match the area of the first Betti curve. However, only hyperbolic spaces with small number of dimensions could also simultaneously match the area of the second Betti curve (Fig. 5B, inset). The 3D hyperbolic space produced the best fit, with larger dimensions yielding increasing deviations. Hyperbolic spaces with dimensions nine and above could be excluded with $P < 0.034$. The third Betti curve was essentially zero and is not shown here. One may notice that the first and second Betti curves were not as regular as in the case of odorants and contained multiple peaks. It turns out that the biphasic nature of the Betti curve could be explained by the nonuniform distribution of points across the two angles (Fig. 5C). Unlike in the case of olfactory stimulus spaces that are sampled approximately uniformly, here, the distribution of points obtained using MDS is not uniform and clusters in one-half of the space. Sampling points from this embedding yields biphasic Betti curves that match those derived from perceptual data (Fig. 5D).

Specifically, P values for L1 differences between Betti curves derived from data and MDS fits were $P = 0.32$ (hyperbolic, Betti 1), $P = 0.20$ (hyperbolic, Betti 2), $P = 0$ (Euclidean, Betti 1), and $P = 0.06$(Euclidean, Betti 2) (cf. inset in Fig. 5D). The MDS distances also better correlated with perceptual distances when we carried out MDS in the hyperbolic space compared to Euclidean space (Fig. S6).

1.4 Discussion

Our results highlight the importance of hyperbolic curved geometry for understanding how natural odors are represented in the nervous system. Overall, we find that both the statistics of natural odor mixtures and human odor perception can be mapped onto hyperbolic spaces. In the natural environment, hierarchical biochemical networks produce odor components. Hierarchical networks can often be approximated by trees and, therefore, by hyperbolic spaces ([7][21]). We find that most natural odor components fall near the boundary of the observed hyperbolic space, corresponding to leaves of the trees (Fig. 1). At the perceptual level, we also found hyperbolic organization. However, in this case, the odors selected for the Dravnieks database did not sample the human perceptual space uniformly (Fig. 5). Hyperbolic perceptual organization is likely to be general across different sensory modalities. There are two reasons for this. First, neural networks that give rise to perception are hierarchically organized, and as we have seen in Fig. 1, this can lead to hyperbolic geometry. Second, individual neurons have limited response ranges. Because of response saturation, small changes in neural responses near their limit correspond to exponentially large changes in the input values. This compressive mapping [22] is similar to the Poincare disk representation of the hyperbolic space [7]. There is evidence that visual, haptic, and auditory perceptual spaces are all hyperbolic ([23][24][25][26]). Adding olfactory perception to this list could help explain why humans can map odors to auditory pitch ([27][28]) and to colors [29]. Noteworthy is the low dimensionality of both the physical odor space and perceptual odor spaces. In both cases, the curved space contains approximately three dimensions despite the fact that the data vary in > 50 dimensions associated with different samples of natural odor mixtures and according > 100 perceptual descriptors. The low dimensionality of the environmental odor space could be a general property of natural odors because it occurred for odors as diverse as fruit and mammalian urine odors. Note that all four natural odor data sets were described by the same 3D hyperbolic space with exactly the same radius (equivalent to curvature of the space). This property could make it easier to represent data from different data sets within the same space. For example, odors from strawberry and tomato could be represented jointly within a single 3D space (Fig. 4). We could not combine data from other data sets because, for example, there were no overlapping components between fruit odors and mouse urine data sets. It is possible that representing all possible natural odors will increase the dimensionality of the overall space. Another possibility is that introducing odors from different sources will "fill in" the inner part of the hyperbolic space. The natural odors considered here mapped onto a surface of a hyperbolic space. Odors produced by biochemical pathways of different complexity are likely to map to surfaces with a different radius, filling in the space. This possibility is especially interesting because it would provide a link to the filled 3D hyperbolic space that we find for perceptual data, which was obtained using diverse classes of odors. At the same time, the perceptual odor mapping reveals that odors tested so far concentrate on one side of the space (Fig. 5) ([9][19]), whereas natural odor components cover their respective space rather uniformly (Fig. 3). These analyses thus suggest perceptual coordinates that are yet to be explored. The match in dimensionality between the environmental and perceptual spaces would not have been expected a priori. The matching dimensionality between the input and perceptual spaces can help avoid nonlinear distortions that would necessarily arise when mapping two nonlinear spaces of different dimensionality. These distortions are known to exist in vision where we perceive distances in a compressed way: The moon appears disproportionately closer to us than would be based on the actual Euclidean distance ([23][24]). We also plot equidistant and parallel lines differently, which is one of the key signatures of the hyperbolic space. Similar distortions arise in the haptic space ([25]). The matching geometry between the input and perceptual spaces in olfaction may therefore serve to minimize these distortions in odor perception. Overall, the ability of the perceptual system to resolve points in the low-dimensional odor space would depend on the number and tuning properties of sensory receptors ([30][31][32][33][34]).

1.5 Materials and Methods

A clique topology method for finding geometric spaces consistent with correlations in the data

We followed procedures from [11] to generate Betti curves for samples taken from spaces with different geometries. The method effectively converts the correlation matrix to its rank-ordered version. This renders the algorithm's results invariant under monotonic transformation of values, for example, due to nonlinearities introduced at the measurement stage. However, this property can also be used to assign a distance between points based on the correlation in the activity of two units in a network [11] or, as in our case, between two odors across different samples. All monotonic functions will yield the same result. We chose $D_{ij} = -|C_{ij}|$, where D_{ij} is the assigned distance between odors, and $|Cij|$ is the absolute value of the correlation coefficient of odor concentrations among a set of points. This definition ensures that stronger correlations (in absolute value) corresponded to tighter connections and smaller geometric distances, as in [35]. The first three Betti curves turn out to be quite sensitive measures of the distance matrices and can be used to find underlying geometries consistent with the data [11]. In addition to random spaces, we screened two kinds of geometric structures: Euclidean spaces of different dimensionality and hyperbolic space [we used the hyperbolic ball model [7] with curvature z = 1] with different parameters. In each space, we uniformly sampled points (the same number as the number of odors in each of the data sets) based on the metric of the space. In a d-dimensional Euclidean space, the points were uniformly distributed in a d-cube with Euclidean distance. For a d-dimensional hyperbolic ball model, we used partial space by setting the minimal radius R_{min} and maximal radius R_{max} for the ball. This choice of the model was motivated by the fact that hyperbolic space approximates hierarchical tree-like networks [7], with odors reflecting leaves—the neighborhood of the surface. We sampled angular coordinates of points uniformly and sampled radial coordinate r within $[R_{min}, R_{max}]$ following the distribution:

$$\rho(r) \sim sinh^{d-1} r \tag{S1}$$

The distance between two points was derived from hyperbolic law of cosines:

$$cosh \zeta x = cosh(\zeta r) cosh(\zeta r') - sinh(\zeta r) sinh(\zeta r') cos \Delta\theta \tag{S2}$$

where ζ is the curvature set as 1 in our model, r and r' are the radial distances of the two points, and $\Delta\theta$ is the angle between them. Considering that noise may exist in the monotonous correspondence between the underlying topological distance and correlation strength of odors and that the amount of noise may differ between data sets due to differences in sample collection procedures, we added multiplicative Gaussian noise to the distance matrices for both Euclidean model and hyperbolic model before plotting the Betti curves. Together, we have the topological distance matrices of the sampled points in geometric spaces:

$$D = D_{geo} \cdot (1 + \varepsilon \cdot N(0,1)) \tag{S3}$$

where D_{geo} is the geometric distances, and ε is the noise level. In summary, the space geometry affects Betti curves through the distribution of sampled point density (Eq. S1) and distance measures (Eq. S2). Multiplicative noise (Eq. S3) also affects Betti curves. The optimal parameter values for the $3D$ hyperbolic model were $R_{max} = 7$ and $R_{min} = 0.9 R_{max}$ for all four odor data sets, while optimal noise values were $\varepsilon = 0.045$ (mouse), $\varepsilon = 0.050$ (strawberry), $\varepsilon = 0.050$ (blueberry), and $\varepsilon = 0.040$ (tomato). The optimized parameters for the Euclidean space were as follows: mouse data set, dimension $d = 8$, $\varepsilon = 0.05$; strawberry data set, dimension $d = 10$, $\varepsilon = 0.05$; blueberry data set, dimension $d = 10$, $\varepsilon = 0.09$; tomato data set, dimension

$d = 8$; $\varepsilon = 0.09$. For the perceptual dataset ([20])) (127 mono-molecular odors, 146 descriptions for each), the topological pairwise distances of odors k and n was defined as $\sum_{i=1}^{146}(v_i^k - v_i^n)$ where v_i^k denote human descriptions for k_{th} odor. We use the differences between descriptions across odors, because in this case the absolute value of the descriptor matter, unlike in the case of odors where correlations were a more appropriate measure. When fitting the data using geometric models, no noise was added to distances in models. We also tested the sensitivity of the Betti curves to noise in pairwise perceptual distances between odors. This was done by computing perceptual distances based on randomly selected subset of 120 out of the total 146 descriptors. The variability in the resultant distance values was proportional to the mean distance (Fig. S7). Importantly, the relative error in the integrated Betti values across these samples was the same as the relative error of the distances themselves (Fig. S7, inset on the right). Thus, although the Betti curve construction evaluates data structure globally, it is not driven by variability in larger distances. In the case of the perceptual dataset, we found that the full hyperbolic space better described the data rather than a shell, and therefore the minimal radius was set to zero. We optimized maximal radius of the hyperbolic model, which is a measure of its curvature, to fit the integrated Betti value of the first Betti curve. The optimal Rmax were as follows: 1.6 (3D), 1.9 (4D), 1.8 (5D), 1.7 (6D), 1.9 (8D), and 3.0 (9D). We used these values to compute the second Betti curve and determine how well it could account for the second integrated Betti value. All reported P values for comparison with experimentally gener- ated Betti curves were obtained by creating, for each candidate geometry, 300 statistically equivalent models. Points for each model were selected randomly according to the density specific to that geometry (uniformly within the unit cube for Euclidean spaces and according to Eq. S1 for hyperbolic spaces). The number of points was matched to the number of points in the corresponding experimental data set. On the basis of this simulated point distribution, we computed 300 different Betti curves. These curves were then used to generate a distribution of integrated Betti values or compute the L1 distance of these curves from the mean Betti curve of this model. The reported P values reflect two- tailed percentiles for where experimental Betti curves fall within the model-generated distributions.

We report P values as < 0.003 when none of the samples generated values further from the mean than the observed data point.

Non-metric MDS embedding of odors on the surface of a 3D hyperbolic sphere

The non-metric MDS algorithm embeds a set of points within a pre-specified space while attempting to preserve rank-ordered distances be- tween points. We modified the Euclidean-based, non-metric MDS algorithm in MATLAB version 2017a by replacing the Euclidean distance with hyperbolic distance in Eq. S2. The initial positions of points were uniformly sampled in the optimal 3D hyperbolic space determined in Fig. 1. The radial coordinates were fixed because their range was small and points were approximately positioned on the surface of a sphere. The algorithm updated the angular coordinates to minimize the mis- match in the rank order of distances. The iterations ended when this error fell below a threshold of 0.001. Because the MDS algorithm can return the arbitrary rotation of the space, in Fig. 3, we used the Procrustes algorithm to align the positions of odors between the strawberry and tomato data sets, using the strawberry data set that had the most odors as an anchor. The Procrustes process was carried out through the Procrustes function in MATLAB, and the scale component and translation component were set to 1 and 0, respectively.

1.6 SUPPLEMENTARY MATERIALS

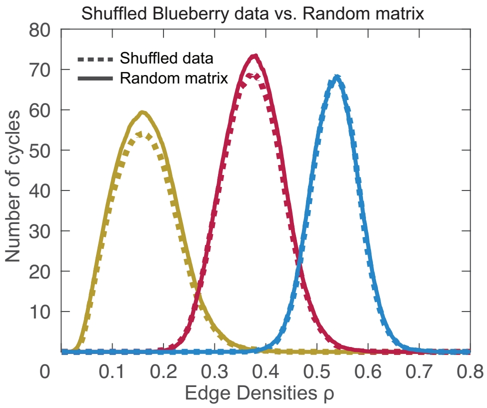

Figure S1. No indications of hyperbolic geometry in shuffled odor data sets. Betti curves from shuffled blueberry data can be accounted for by random sampling and is not consistent with the Hyperbolic or Euclidean models. The correlation matrix between odors was computed by taking measurements of odor pairs from separate, randomly selected fruit samples. This removes correlations in the fluctuations of component concentrations between odors. Computing Betti curves from such shuffled correlation matrices produces Betti curves that are fully consistent with random matrices ($P = 0.4, 0.7, 0.9$ for Betti curves one, two and three, respectively) and are not consistent ($P < 0.02$) with either Euclidean or hyperbolic spaces with small noise amounts as necessary to account for real (unshuffled) data.

Table S1. Statistical tests (P values) for consistency with hyperbolic models based on integrated Betti values.

p values		Dimension								
		3	4	5	6	8	10	20	50	100
Mouse urine	Betti 1	0.973	0.740	0.900	0.967	0.933	0.953	0.887	0.807	0.807
	Betti 2	0.453	0.853	0.980	1.000	0.920	0.773	0.540	0.573	0.593
	Betti 3	0.733	0.313	0.313	0.347	0.440	0.700	0.800	0.680	0.533
Strawberry	Betti 1	0.967	1.000	0.953	0.747	0.860	0.860	0.967	0.880	1.000
	Betti 2	0.893	0.613	0.713	0.420	1.000	0.813	0.707	0.647	0.680
	Betti 3	0.193	0.033	0.080	0.087	0.273	0.287	1.000	0.940	0.967
Blueberry	Betti 1	0.247	0.920	0.860	0.993	0.947	0.813	0.880	0.847	0.833
	Betti 2	0.260	0.173	0.167	0.127	0.080	0.093	0.040	0.027	0.013
	Betti 3	0.580	0.887	0.880	1.000	0.820	0.820	0.347	0.453	0.387
Tomato	Betti 1	0.747	0.973	1.000	0.880	0.913	0.840	0.867	0.947	1.000
	Betti 2	0.253	0.947	0.900	0.733	0.653	0.400	0.193	0.313	0.333
	Betti 3	0.207	1.000	0.873	0.633	0.533	0.187	0.007	0.033	0.000

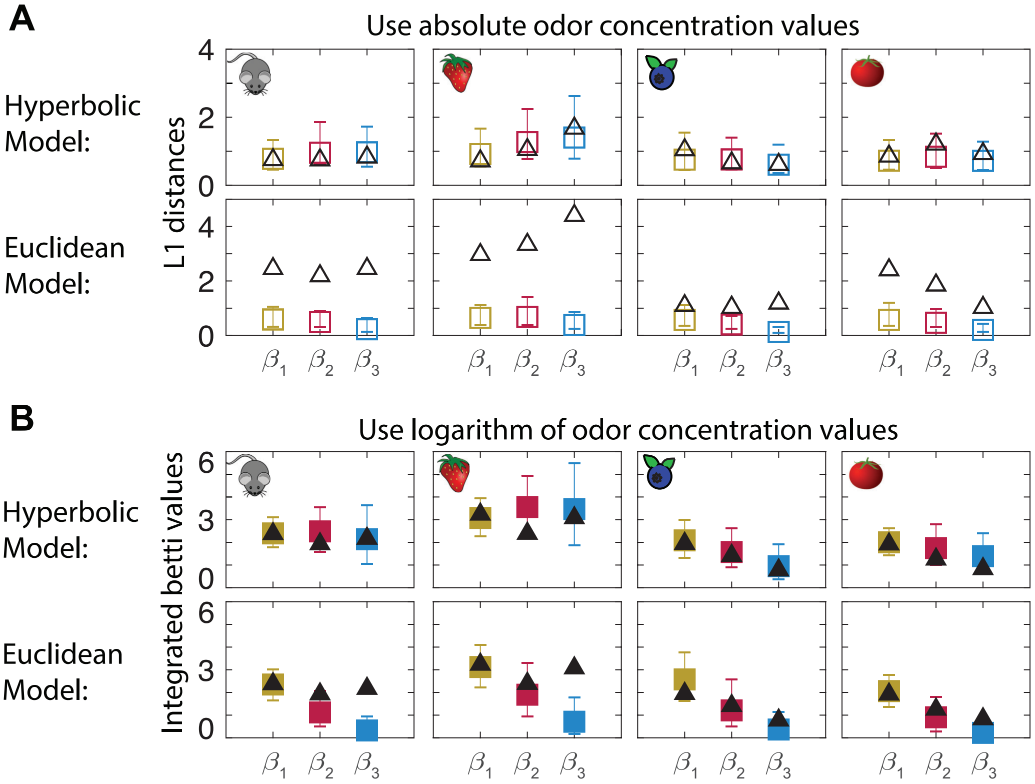

Figure S2. Alternative ways of evaluating differences between Betti curves also support hyperbolic geometry of natural odor spaces. Error bar plots of Betti curves statistics of both geometric models using L1 distances (A) or logarithm of concentrations and integrated Betti values (B). The same geometric parameters were used as in Fig. 2. (A) The gray triangles showed the L1 distance between Betti curves of data and the mean of 300 geometric models, the error bar plots showed the statistics of the L1 distances between Betti curves of all of the 300 models and model mean. The error bar showed the 95% confidence intervals (2.5% $\sim$ 97.5%). (B) Use the logarithms of odors concentrations to generate Betti curves. The gray triangles and colored error bar plots show the statistics of integrated Betti values in both geometric models.

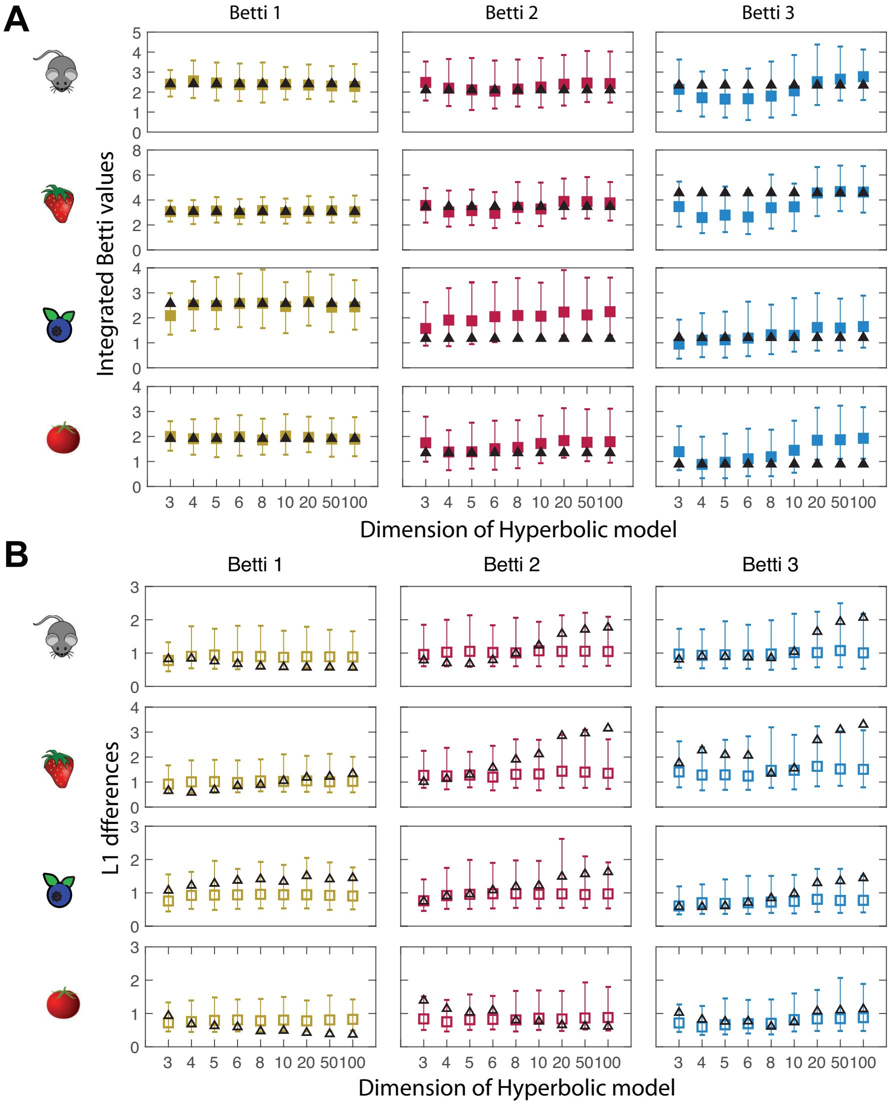

Figure S3. Error bar plots of Betti curves statistics for the hyperbolic model of different dimensions. The parameters were determined by fitting the first integrated Betti value of the models to data (first columns), then the same parameters were used to evaluate the correspondence with the second and third Betti curves. (A) Error bar plots of integrated Betti values of data (gray triangles) and 300 model repeats (colored bars) with dimension ranging from 3D to 100D. (B) Error bar plots of L1 difference of Betti curves between data and model mean (gray triangles), and the distances between all 300 models and model mean (colored bars). The parameters were the same as in Fig. 2 and Fig. S2.

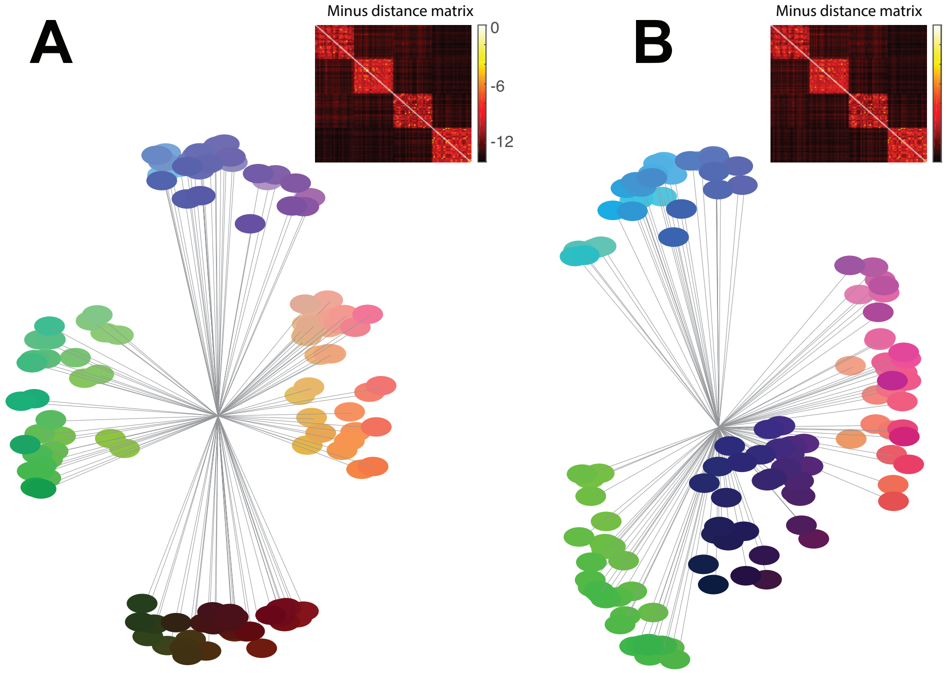

Figure S4. Test of the non-metric multidimensional scaling algorithm in the hyperbolic space on synthetic data. (A) 120 sampling points were generated near the surface of 3D hyperbolic sphere forming four clusters. The radii were distributed uniformly within 0.9 of the sphere radius, the same distribution we used to model natural odor mixtures. Inset in the top right shows the matrix of pairwise distances that indicates these four clusters. (B) Non-metric multidimensional scaling can be used to embed points on the surface of a 3D hyperbolic sphere. The embedded points also form four clusters, albeit at different orientation in the space. The distance matrix (inset) is also reproduced.

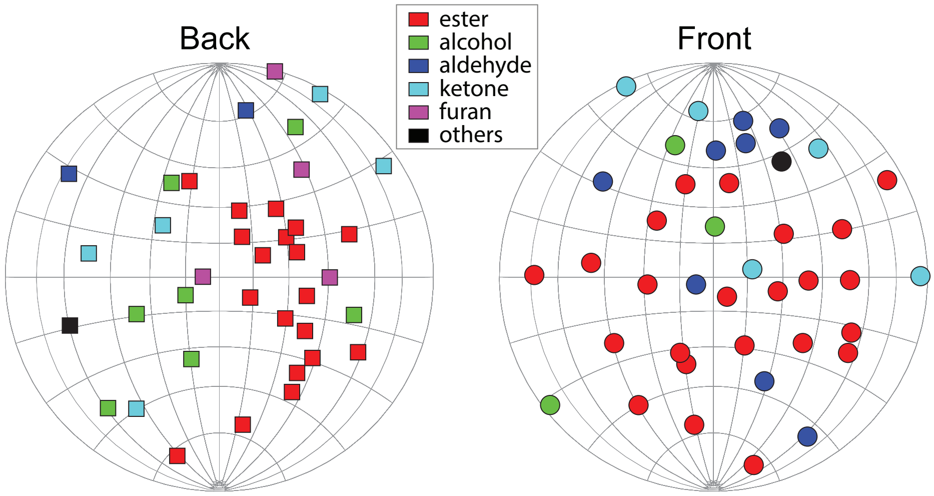

Figure S5. Odors within the identified space do not cluster by functional group. The odors shown are from the strawberry data set. Circles and squares show the front and back side of the sphere, respectively.

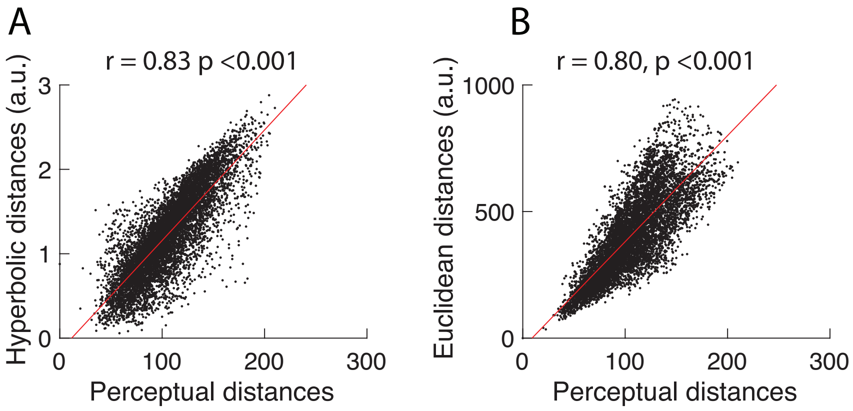

Figure S6. Comparison between embedded geometric distances and reported perceptual distances. Non-metric multidimensional scaling was used to embed the odors into 3D Hyperbolic space (A) or 3D Euclidean space (B).

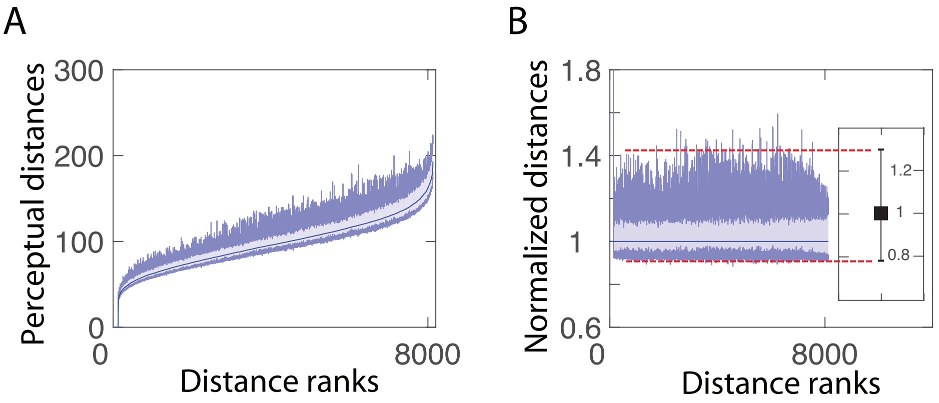

Figure S7. Analysis of sensitivity of integrated Betti value to noise in the input distances. (A) The error bars show 95% confidence intervals for the distributions of Betti value computed based on 300 different partial samples of perceptual descriptors (120 out of 146 perceptual descriptors). Odor distances are rank ordered based on the medium distance across 300 samples. The blue line near the center shows the medium pairwise distances. (B) Pairwise distances normalized by their medium. The error bar plots show the 95% percent confidence intervals, blue line is the medium. Variability in normalized distances no longer depends on distance and matches the variability in the first integrated Betti value normalized by its medium (inset).

Table S2. Statistical tests (P values) for consistency with hyperbolic models based on L1 distances between Betti curves.

p values		Dimension								
		3	4	5	6	8	10	20	50	100
Mouse urine	Betti 1	0.873	0.800	0.433	0.327	0.187	0.107	0.093	0.080	0.113
	Betti 2	0.347	0.147	0.173	0.413	0.967	0.613	0.287	0.180	0.147
	Betti 3	0.553	0.913	0.853	0.833	0.720	1.000	0.280	0.113	0.080
Strawberry	Betti 1	0.173	0.060	0.173	0.640	0.613	1.000	0.713	0.500	0.427
	Betti 2	0.287	0.767	1.000	0.413	0.327	0.233	0.060	0.067	0.013
	Betti 3	0.513	0.073	0.167	0.133	0.793	1.000	0.173	0.040	0.027
Blueberry	Betti 1	0.313	0.353	0.367	0.267	0.260	0.393	0.220	0.200	0.180
	Betti 2	0.967	0.947	1.000	0.700	0.587	0.493	0.327	0.213	0.153
	Betti 3	0.920	0.607	0.747	1.000	0.660	0.460	0.213	0.120	0.080
Tomato	Betti 1	0.373	0.713	0.373	0.300	0.067	0.060	0.007	0.007	0.007
	Betti 2	0.067	0.140	0.387	0.387	0.960	0.693	0.340	0.193	0.193
	Betti 3	0.220	0.340	0.607	0.733	0.673	0.807	0.493	0.453	0.460

Table S3. Statistical tests (P values) for evaluating consistency of experimental Betti curves with respect to 3D hyperbolic model or optimal optimal Euclidean model.

p values		Integrated betti values		L1 differences	
		3D Hyperbolic	Optimal Euclidean	3D Hyperbolic	Optimal Euclidean
Mouse urine	Betti 1	0.973	0.907	0.873	0.000
	Betti 2	0.453	0.047	0.347	0.000
	Betti 3	0.733	0.000	0.553	0.000
Strawberry	Betti 1	0.967	0.893	0.173	0.000
	Betti 2	0.893	0.020	0.287	0.000
	Betti 3	0.193	0.000	0.513	0.000
Blueberry	Betti 1	0.247	0.993	0.313	0.060
	Betti 2	0.260	0.940	0.967	0.000
	Betti 3	0.580	0.027	0.920	0.000
Tomato	Betti 1	0.747	0.620	0.373	0.000
	Betti 2	0.253	0.333	0.067	0.000
	Betti 3	0.207	0.000	0.220	0.000

Table S4. Statistical tests (P values) for evaluating consistency of Betti curves computed based on logarithm of odor concentrations with respect to hyperbolic model. Consistency evaluated based on integrated Betti values

p values		Dimension								
		3	4	5	6	8	10	20	50	100
Mouse urine	Betti 1	0.947	0.693	0.840	1.000	1.000	0.993	1.000	0.867	0.840
	Betti 2	0.293	0.547	0.713	0.833	0.680	0.507	0.347	0.380	0.393
	Betti 3	0.960	0.473	0.453	0.427	0.627	0.853	0.547	0.513	0.360
Strawberry	Betti 1	0.707	0.740	0.773	0.553	0.867	0.633	0.807	0.780	0.740
	Betti 2	0.060	0.327	0.273	0.433	0.147	0.220	0.027	0.027	0.073
	Betti 3	0.680	0.640	0.713	0.633	0.767	0.653	0.107	0.047	0.080
Blueberry	Betti 1	0.713	0.307	0.287	0.293	0.233	0.287	0.200	0.367	0.300
	Betti 2	0.687	0.460	0.353	0.340	0.240	0.307	0.113	0.180	0.113
	Betti 3	0.607	0.453	0.420	0.333	0.233	0.167	0.087	0.087	0.040
Tomato	Betti 1	0.753	0.973	1.000	0.893	0.907	0.847	0.873	0.933	1.000
	Betti 2	0.207	0.793	0.753	0.547	0.540	0.320	0.120	0.193	0.213
	Betti 3	0.167	0.900	0.733	0.540	0.427	0.140	0.007	0.020	0.000

Table S5. P values of hyperbolic and Euclidean model using integrated Betti values for perceptual data set.

p values		Hyper 3D	Hyper 4D	Hyper 5D	Hyper 6D	Hyper 8D	Hyper 9D	Euc 2D	Euc 3D	Euc 4D
Odor atlas	Betti 1	0.920	0.860	0.893	0.987	0.767	0.680	0.000	0.120	0.000
	Betti 2	0.187	0.227	0.273	0.133	0.053	0.033	0.000	0.013	0.000

Chapter 1, in full, is a reprint of the material as it appears in Hyperbolic geometry of the olfactory space. *Science advances*, 2018; Yuansheng Zhou, Brian H Smith, Tatyana O Sharpee. The dissertation author was the primary investigator and author of this paper.

Chapter 2

Hyperbolic geometry of gene expression

2.1 Abstract

Patterns of gene expressions play a key role in determining cell state. Although correlations in gene expressions have been well documented, most of the current methods treat them as independent variables. One way to take into account gene correlations is to find a low-dimensional curved geometry that describes variation in the data. Here we develop such a method and find that gene expression across multiple cell types exhibit a low dimensional hyperbolic structure. When more genes are taken into account, hyperbolic effects become stronger but representation remains low-dimensional. The size of the hyperbolic map, which indicates the hierarchical depth of the data, was the largest for human cells, the smallest for mouse embryonic cells and intermediate in differentiated cells from different mouse organs. We also describe how hyperbolic metric can be incorporated into the t-SNE method to lead to improved visualizations compared to leading methods.

2.2 Introduction

One of the great challenges of modern biology is to understand how the genotype of an organism impacts its phenotype, such as disease risk. The difficulty of this problem stems from the complexity of this relationship where thousands of genes can affect a phenotype of interest through nonlinear interactions [36, 37, 38]. In the past 15 years, genome-wide associations studies have demonstrated that a range of traits, including those that related to metabolic and mental health disorders, are potentially linked to thousands of genes, with each gene explaining only a small fraction of the expected heritability [37]. At the same time, correlations between genes are widespread [39]. These observations raise the possibility that genetic variation and their expression can be described by a low-dimensional geometry. Identifying this geometry would make it easier to find relevant gene combinations and how they impact a given trait.

Traditional approaches to finding low-dimensional spaces, such as the principal component analysis (PCA), assume that the space is "flat" (i.e. has zero curvature) and evaluate distances between points according to Euclidean metric. Recently hyperbolic spaces have attracted a lot of attention both for the analysis of biological data [40, 41, 42] as well as in computer science [43, 44, 45, 46, 47, 48, 49]. The reason for this interest is that hyperbolic metric approximates the exponential expansion of possible states of the system described by a hierarchical tree-like process [7]. Hierarchical representations, such as phylogenetic trees and clustering clades have long been used to characterize differences between cells [50], proteins [51], the activity of metabolic networks within cells [52, 53] and human brain functional networks [54]. This suggests that hyperbolic metric should be considered as one of the possibilities when searching for the low-dimensional geometry in biological data. At the same time, any hyperbolic geometry (which has negative curvature) can locally be approximated using Euclidean geometry (which has zero curvature). Therefore, in this work we focus on comparing the signatures of Euclidean and hyperbolic geometry. For completeness, we also include results from the spherical geometry that has positive curvature and represents the last of three possible geometries with constant curvature.

In this work we pursue two goals. The first goal is to develop a quantitative test for distinguishing the curvature of the underlying low-dimensional geometry. We show that this can be achieved by performing non-metric multi-dimensional scaling using both Euclidean and hyperbolic metric and comparing the results. Our second goal is to develop visualization tools for data that exhibit a low-dimensional hyperbolic geometry. Many of the current state-of-the-art visualization tools, such as k-means clustering [55], local linear embedding [56], t-distributed Stochastic Neighbor Embedding (t-SNE) [57, 58], and Uniform Manifold Approximation and Projection (UMAP) [59], all use Euclidean metric. We propose a method for incorporating hyperbolic metric into the t-SNE method and show that this leads to improved visualization across a range of datasets.

To demonstrate the utility of both the diagnostic method and the hyperbolic t-SNE (h-SNE), we apply these methods to a range of gene expression data sets from mouse and human. These datasets uniformly show that gene expression data across different cell types exhibits a low-dimensional hyperbolic geometry. The curvature of this space, which is related to the branching ratio of the corresponding tree-like process, was systematically higher in differentiated cell types compared to embryonic cells, and took even larger values for brain cells. These results demonstrate that gene expression data can be effectively described using a small number of coordinates under hyperbolic metric. Visualizations using hyperbolic metric consistently showed more accurate representations, both in terms of local and large-scale structure, including more consistent estimates of developmental states in datasets where pseudo-time trajectories could be constructed [41].

2.3 Results

2.3.1 Non-metric MDS outperforms metric-MDS in geometry detection

Multi-dimensional scaling (MDS) has been widely used to embed a set of data points into a geometric space in a way that attempts to best preserve the distances between points in the original space. Metric MDS tries to make the embedding distances proportional to the input distances, while non-metric MDS preserves the only ordinal values, allowing a monotonic nonlinear transformation between the distances. Both metric and non-metric MDS in high dimensional Euclidean space have been well studied during the past few decades. However, the MDS in the hyperbolic space has not been fully developed yet. Several metric MDS algorithms

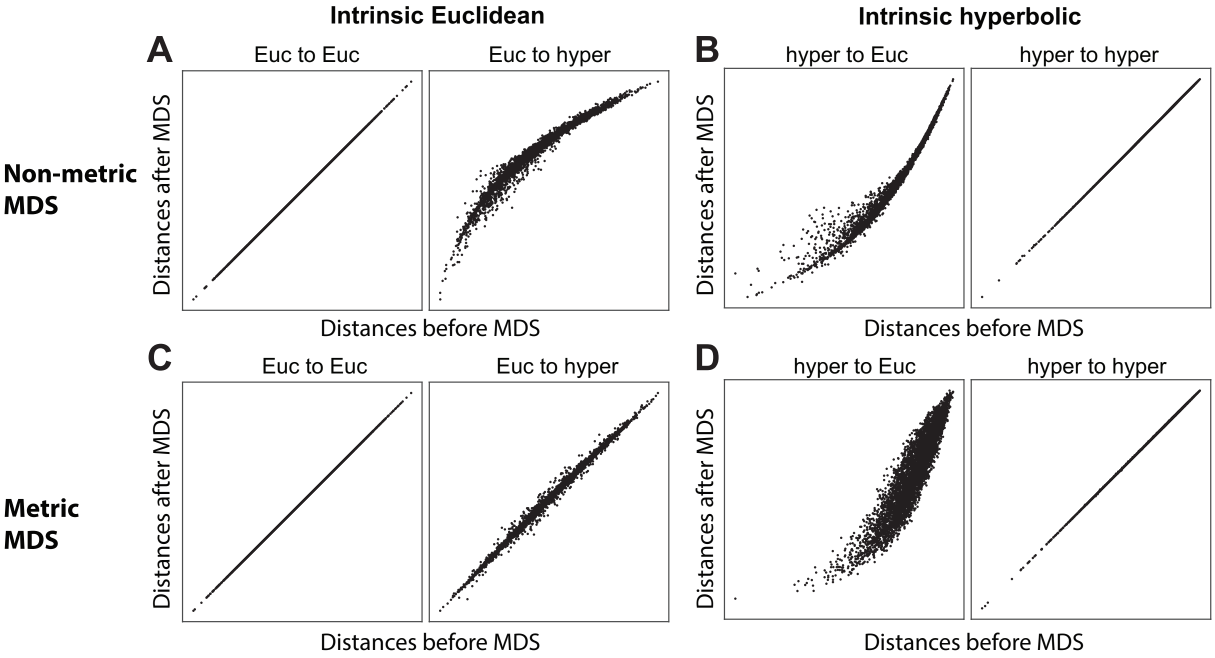

Figure 6. Shepard diagrams for metric MDS and non-metric MDS applied to synthetic geometric data with either Euclidean or hyperbolic native geometry. These simulations were produced by (1) randomly sampling 100 points in 5D Euclidean (A,C) and hyperbolic space (B,D), (2) computing geometric distances using the corresponding distance metrics, and (3) using either non-metric MDS (A-B) or metric MDS (C-D) to embed the points to 5D Euclidean and hyperbolic space. The radius of the hyperbolic space (measured in units of inverse curvature) was 3.0 for both the sampling and embedding spaces. The convexity of Shepard diagrams reflects the difference in geometry between the embedding and native spaces, it is positive when hyperbolic data is embedded in Euclidean space, and negative when Euclidean data is embedded into the hyperbolic space. These differences are less distinct in the case of metric MDS (bottom row).

have been proposed recently for embedding data into hyperbolic space, offering advantages over Euclidean visualizations in terms of distance preservation, space capacity, trajectory inference and unseen data prediction [60, 41, 43, 44, 45, 46, 47, 48, 49, 42], etc. However, we find that metric MDS does not correctly distinguish between Euclidean and hyperbolic geometry of input data, but non-metric MDS does (Figure 6). The reason for this is that non-metric MDS matches the ranking order instead of exact values of the data distances. The resulting nonlinear distortions in embedding distances can be used as indicators for a geometry mismatch between data and embedding points. When using non-metric MDS, we illustrate that as soon as there is a mismatch between native and embedding geometry, a nonlinear distortion appears in the scatter plots of embedding distances versus input data distances (Figure 6A-B). These scatter plots are known as Shepard diagrams [61]. When Euclidean data is embedded into a hyperbolic space, the Shepard diagram has negative convexity (Figure 6A). When hyperbolic data is embedded into Euclidean space, the Shepard diagram has a positive convexity (Figure 6B). Thus, the convexity of the Shepard diagram can indicate the difference in geometric properties between the embedding and native spaces, and in particular could indicate the difference in curvature of geometry. When using the metric MDS, the Shepard diagram shows increased spread (Figure 6D) but does not yield a nonlinear relationship upon embedding Euclidean data to hyperbolic space (Figure 6C). The reason is that Euclidean distances can be fully embedded into the faster-expanding hyperbolic space masking the distortion of distances, and this does not happen in the non-metric MDS [61]. In what follows we apply non-metric MDS to synthetic and several real gene expression datasets to detect their hidden geometry, and we refer non-metric MDS as simply MDS for brevity.

2.3.2 Synthetic geometric data

When cells are characterized according to the expression of thousands of genes, the number of genes represents the nominal dimension of the representation space. However, the real dimension of the gene expression space might be much lower. Furthermore, the true geometry of the hidden space is not necessarily Euclidean. Therefore, in this section we analyze the

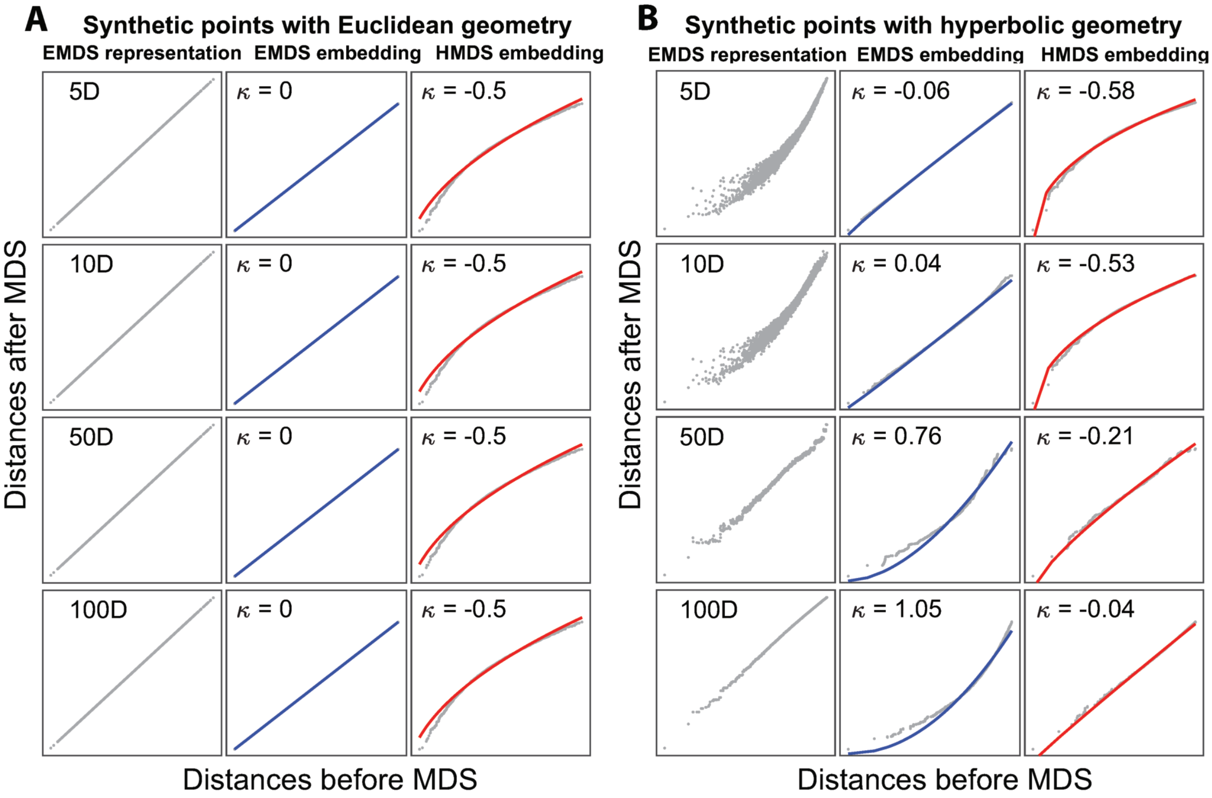

Figure 7. Illustration of the diagnostic approach base don NDS for geometry detection on synthetic data. (A) Randomly sampled 100 points from 5D Euclidean space are embedded into 5, 10, 50 and 100 dimensional Euclidean spaces (left), followed by subsequent embeddings into 5D Euclidean space (middle) or hyperbolic space (right). The solid lines represent the fits using Equation (S1), the Shepard diagram curvatures κ are shown at the top of each panel. (B) Same analysis for 100 sampled points from a 5D hyperbolic space with $R_{\text{data}} = 3.0$. The hyperbolic radii used in HMDS are $R_{\text{model}} = 3.0$ in both (A) and (B).

signatures of low dimensional geometry of constant curvature (either Euclidean, spherical, or hyperbolic) in the situation where each data point is described with respect to large number of variables. In the synthetic examples below, the points are first sampled from a low dimensional geometry and then embedded into a high dimensional Euclidean space. This step is included to mimic analysis of experimental data, where each data point is evaluated according to a large number of measurements. After this, the data points are embedded into spaces of different curvatures to determine indicators through which the properties of the original low-dimensional space can become apparent. In the examples below, we focus primarily on hyperbolic and Euclidean geometries, because hyperbolic geometry describes hierarchically organized data, whereas Euclidean metric is often the only feasible geometric metric for computing distances of high dimensional vectors. Comparison with the results for spherical spaces is provided in Figure S1.

First, we analyze the case where data has a 5D Euclidean underlying geometry. To simulate this case we randomly sample 100 points from a 5D Euclidean space, and use Euclidean MDS (EMDS) to embed the points to 5D, 10D, 50D and 100D space respectively (Figure 7A, left). This step emulates the representation of real data where each data point is described by a large number of measurements (e.g. transcriptome) according to which each cell is characterized, and the distances between points are measured according to a Euclidean metric. The embeddings with different number of dimensions correspond to cases where measurements are taken with respect to different number of genes. As expected, the distances of synthetic 5D Euclidean points can be preserved without distortion when embedding data to Euclidean spaces of higher dimensions. This is evidenced by the linearity of Shepard diagrams in left column of Figure 7A. Next, we apply EMDS (Figure 7A, middle) and hyperbolic MDS (HMDS) (Figure 7A, right) to the points in the Euclidean representation space, as we did in Figure 6A-B. As one can see in Figure 7A, Euclidean embeddings of these data do not generate distortions in the Shepard diagrams but hyperbolic embeddings yield Shepard diagrams with negative convexity that is largely independent of embedding dimensions. This indicates that the data has an underlying

Euclidean geometry.

To quantitatively characterize Shepard diagrams we fit them using:

$$y = a \cdot (x - x_0)^{\kappa+1}, \tag{S1}$$

where x and y represent distances of points before and after embedding, respectively; parameter $x_0 = \min(x) - \varepsilon$ is the distance offset representing the difference caused by noise from biological variations or experimental measurements, with a small ε introduced to avoid zero input values in the fitting. The parameter κ is key because it characterizes the convexity of Shepard diagrams. The zero $\kappa = 0$ indicates pure linearity and an exact match between the model and data geometries, while $\kappa \neq 0$ indicates convexity and a mismatch between the two geometries. So the sign of κ can indicate the difference in curvature between two spaces. In the current examples, $\kappa = 0$ in EMDS and $\kappa = -0.5$ in HMDS embedding with $R_{\text{model}} = 3$, with no changes as the representation dimension (Figure 7A). These values describe the signatures of data that has intrinsic Euclidean geometry (Figure 6A).

The situation is qualitatively different for the case where the data has a hidden hyperbolic geometry, cf. Figure 7B. Here we sample 100 points from a 5D hyperbolic space with $R_{\text{data}} = 3.0$. The initial embedding of these points into a low-dimensional Euclidean space produces distortions, indicating that using Euclidean metric to evaluate hyperbolic distances between points will not be accurate when using the same dimension for the embedding space. However, Euclidean embeddings into larger dimensional spaces can produce accurate distance representations. For example, in the left column of Figure 7B , accurate distance representation is obtained starting with ~ 50 embedding dimensions. The reason for this is that in a large dimensional Euclidean space points could be distributed along hyperbolic manifolds, approximating the true hyperbolic metric. We next apply MDS to examine the geometry of the representations by embedding them into a low dimensional Euclidean (middle column) or hyperbolic space (right column). With the increase of representation dimension, the convexity parameter κ of the Shepard diagram increases from approximately zero value ($\kappa = -0.06$) to $\kappa = 1.05$ in EMDS and increases from $\kappa = -0.58$ to an approximately zero value ($\kappa = -0.04$) in HMDS (Figure 7B). These signatures (Figure 6B) indicate that hyperbolic property is more fully preserved when points are characterized with respect to more dimensions. These analyses of synthetic data illustrate how a combination of EMDS and HMDS can be used to elucidate the intrinsic geometry starting with the initial Euclidean representation. This method can also be used to detect spherical geometry which has positive curvature. Synthetic results show that spherical geometry has opposite property as hyperbolic geometry: spherical is to Euclidean looks like what Euclidean is to hyperbolic in Shepard diagram (Figure S1).

2.3.3 Geometry of gene expression data

We now apply this method to analyze the intrinsic geometry of gene expression data. We first analyze a discrete gene expression data from Lukk et al. [62]. In the paper, they integrated microarray data from 5372 human samples representing 369 different cell and tissue types, disease states and cell lines, which has a complex global structure. They constructed a global gene expression map by performing principal component analysis and found that the first two principal axes described variation in biological variables corresponding to hematopoietic and malignancy properties. However, the presence and properties of the underlying low-dimensional geometry and how the samples are organized in the space remain to be investigated. Several previous studies showed that gene expressixton was stochastic both at the single cell level and the population level [63, 64, 65], and the expression profiles of samples within the same cluster were dominated by intrinsic noise [63]. This would imply either Euclidean geometry, at least locally, or a lack of geometric structure altogether. On a global scale, biological systems usually show a hierarchical structure which would imply hyperbolic geometry [52, 54]. Therefore, we separately probe the geometry of gene expression data at the local and global scales. To probe local geometry we apply k-means ($k = 50$) method to cluster the whole data and select 100 samples from a single cluster randomly. Similarly to Figure 7, we use increasing subsets of

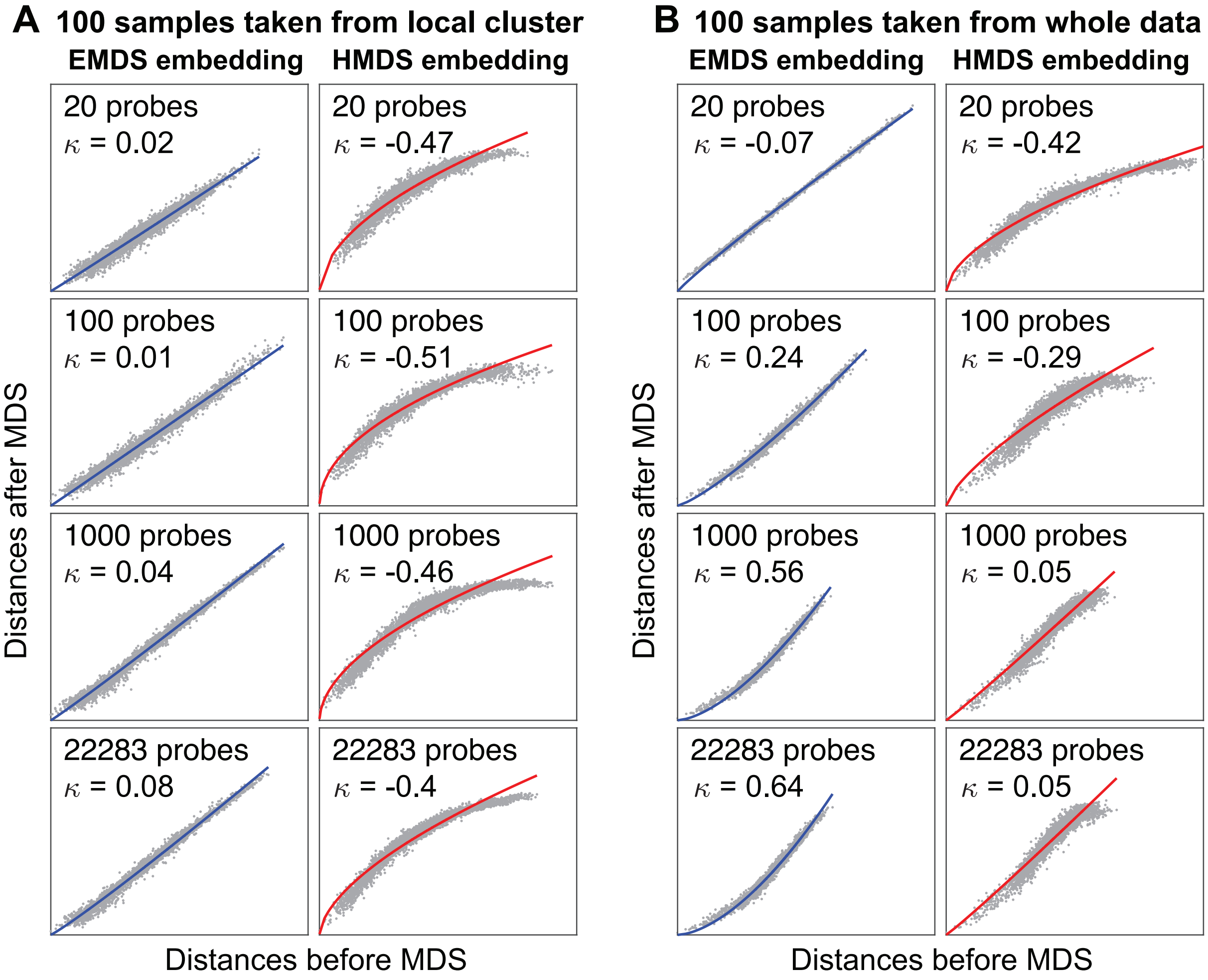

Figure 8. Human gene expression has locally Euclidean and globally hyperbolic hidden geometry. (**A**) MDS embedding results for samples taken from a single k-means cluster, with distances evaluated by Euclidean metric with respect to increasing number of probes. Left and right columns show results of embeddings into 5D Euclidean and hyperbolic space, respectively. (**B**) Same analysis for samples taken randomly from the whole data. The hyperbolic space used in HMDS had radius $R_{\text{model}} = 2.6$ in both (A) and (B).

genes (from 20 probes to all the 22283 probes) to represent samples and then perform EMDS and HMDS embeddings ($R_{\text{model}} = 2.6$) for geometry detection (Figure 8A). Increasing the number of probes with respect to which samples were characterized corresponds to increasing the dimensionality of the initial Euclidean embedding as in Figure 7. We find that this does not significantly change the convexiy of the Shepard diagram in both EMDS and HMDS ($\kappa \approx 0$ in EMDS and $\kappa \leq -0.4$ in HMDS). These results match the fitting in Figure 7A, and indicate that the samples taken from the same cluster have Euclidean structure, even when all the probes are used (Figure 8A). Additional analyses show that the Euclidean structure is indeed caused by the stochastic Gaussian expressions of genes among the samples within a cluster (Figure S2).

Variations in gene expression across samples taken from different clusters, which represent different cell types, tissues and disease states, show more complicated distributions (Figure S2) and have attracted a great deal of attention [66]. To study the geometric structure of expression space globally, we selected 100 samples randomly from the whole population instead of local clusters, and performed the same embeddings as in Figure 8A. Surprisingly we find that, as the number of probes increases, the convexity of the Shepard diagram increases from being approximately zero $\kappa = -0.07$ to $\kappa = 0.64$ in EMDS and from $\kappa = -0.42$ to $\kappa = 0.05$ in HMDS (Figure 8B). These fitting results match the signatures expected for hyperbolic geometry in Figure 7B. It shows that the gene expression space has hyperbolic structure that becomes increasingly more apparent upon including a moderately large number of genes (> 1000 probes) in the measurements.

To test robustness of this conclusion and make full use of the whole data, we repeat the sampling process 300 times both for the local sampling where samples are taken from different single clusters, and for the global sampling where samples are broadly taken from the whole data. The samples are taken with replacement. As expected, for samples taken from local clusters, the median values of convexity $\kappa \approx 0$ in EMDS and $\kappa < 0$ in HMDS (Figure 9A-B) even when all genes are used. These measurements indicate Euclidean structure. For samples taken across the whole population, with increasing number of probes, the median of κ increases to be positive in

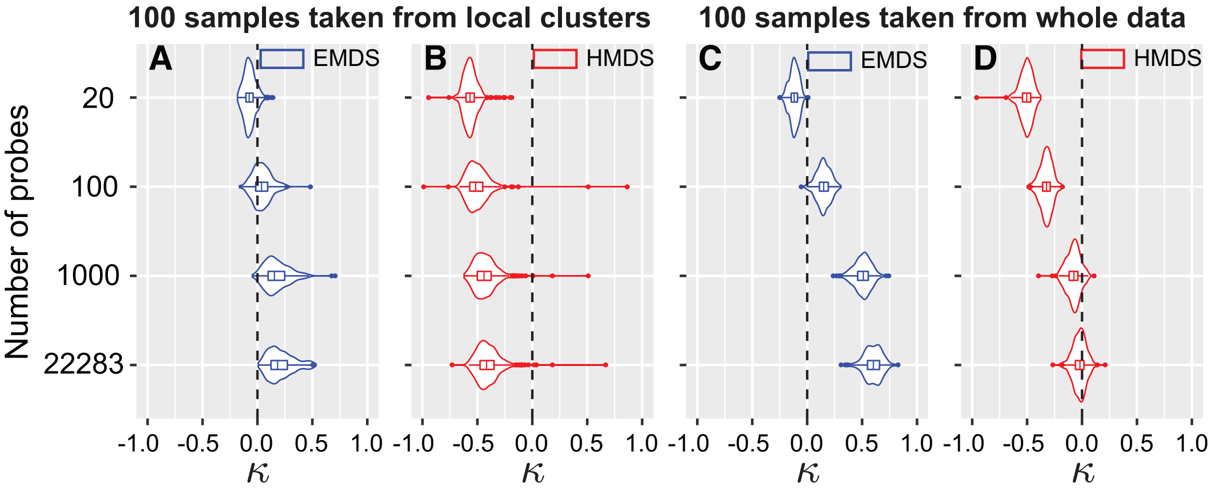

Figure 9. Statistics of convexity parameter of Shepard diagrams across human gene expression data. (A-B) Violin plots show the convexity statistics in 5D EMDS (A) and 5D HMDS (B) across 300 repeated sampling from different k-means clusters, as a function of the number of probes. 100 data points are taken in each sampling. (C-D) Same analysis for samples taken with replacement from the whole data. The black dashed lines show $\kappa = 0$ and signify Euclidean geometry in EMDS (A and C) or hyperbolic geometry with $R_{\text{data}} = R_{\text{model}} = 2.6$ in 5D HMDS (B and D). In each plot, the width of the shape shows the probability density of different values; the central line, the left edge and right edge of the box within the shape represent the median, the 75th and 25th percentiles respectively. The line within the shape extends to the most extreme non-outlier points, the outliers are represented by dots.

EMDS and close to zero in HMDS (Figure 9C-D); these signatures indicate that samples across population have hyperbolic structure when represented by a moderately large number of genes ($\geq$ 1000 probes).

2.3.4 The size of the hyperbolic gene expression map varies systematically across cell types

The HMDS method can also be used to estimate the curvature of the underlying low-dimensional space or equivalently the size of the hyperbolic map measured in units of inverse curvature (Figure S3). The hyperbolic radius R_{data} can be used as an indication of the hierarchical depth of the corresponding tree structure [7]. Above we have shown that global human gene expression data can be embedded without distortion to 5D hyperbolic space with $R_{\text{model}} = 2.6$. This value was obtained by systematically screening across different R values to find those

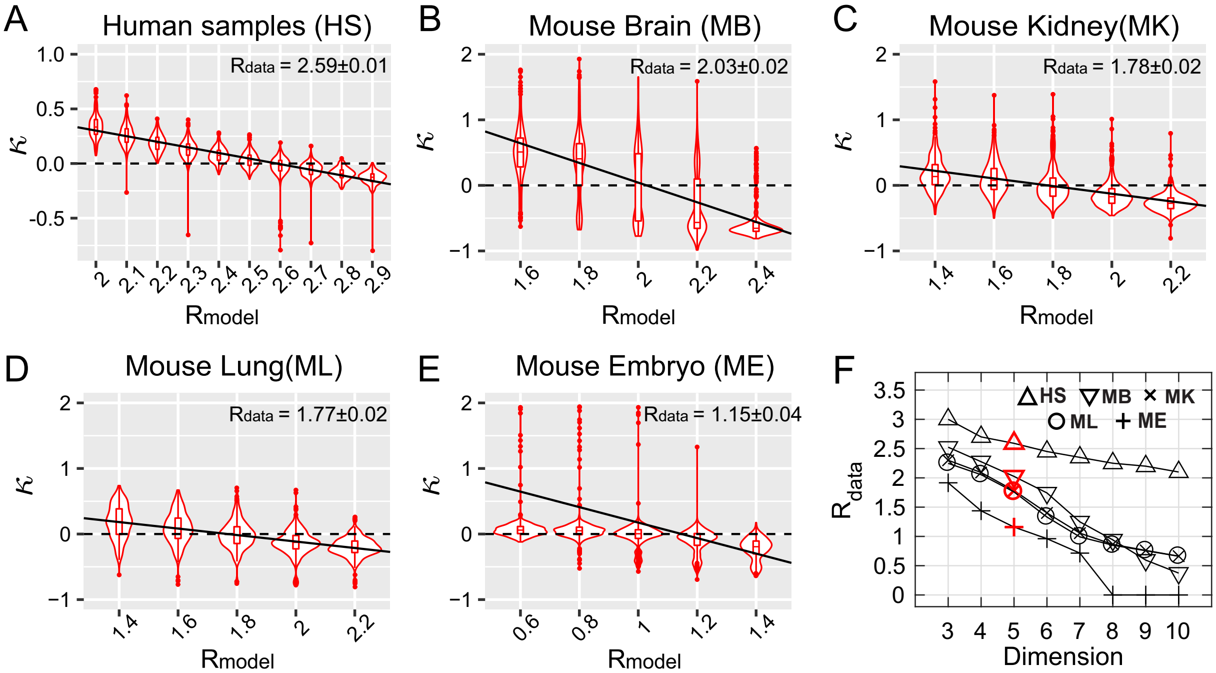

Figure 10. Radius of the hyperbolic space of gene expression varies systematically across cell types. (A-E) The violin plots of convexity of Shepard diagrams κ as a function of the radius of the embedding space R_{model}. (**A**) microarray data from human samples in Lukk et al. [62] dataset. (**B-E**) Microwell-seq data [67] for brain cells (**B**), kidney (**C**), lung (**D**) and embryonic stem cells (**E**). Solid lines show the linear regression; their y-intercepts yield estimates of the radius of the hyperbolic map for each datasets (see printed values with ± 2 standard deviation from 300 sampling). The HMDS embedding dimension is $D = 5$ in (A-E). (**F**) Dependence of the radius of the hyperbolic space as a function of the embedding dimension for datasets in panels (A-E) on human cells (HS), mouse brain (MB), mouse kidney (MK), mouse lung (ML), and mouse embryonic cells (ME). $R_{\text{data}} = 0$ ($D \geq 8$ in ME) means that κ remains to be negative regardless of R_{model}.

best matching R_{data} as indicated by the zero convexity parameter κ obtained by fitting the corresponding Shepard diagram. In Figure 10A we show that the convexity κ decreases with R_{model} and crosses 0 at $R_{\text{model}} \approx 2.6$. Therefore, we conclude that $R_{\text{data}} = 2.6$ in 5D hyperbolic representation. Next, we examine several other gene expression datasets and determine R_{data} for them. Han et al. [67] performed Microwell-seq of cells from multiple mouse organs and generated mouse cell atlas map using the t-SNE method. Here we re-analyze this data to find if it has a nonlinear low-dimensional structure. The microwell-seq data in this dataset are much sparser than the microarray data. Therefore, we first check whether changes in sparseness of measurement data could affect the geometry detection and its parameters. To this end, we re-analyze synthetic data where at the stage of Euclidean high-dimensional embedding, all values are re-set to zero if their values are in the smallest 5%. Even though intermediate embeddings into larger dimensional space have more values that are set to zero, this does not change the estimated convexity values for low-dimensional embeddings using either Euclidean or hyperbolic metric, and the tests correctly identify the presence of a low-dimensional hyperbolic geometry (Figure S4).

With these checks at hand, we proceeded to analyze the microwell-seq data from different mouse organs. Following previous studies [68], the data was pre-processed using the Seurat algorithm [69] and projected onto top 50 principle components (see Methods). Next, we applied 5D HMDS to the processed data from four of the mouse organs – brain, kidney, lung and embryonic stem cells. We find that all these data have an underlying hyperbolic structure (Figure 10B-E). It is worth noting that the hyperbolic radius necessary to describe these data is smaller than that for human samples. Among the four mouse cell types, the largest radius is found for the mouse brain cells with $R_{\text{brain}} = 2.03 \pm 0.02$ (Figure 10B), followed by mouse kidney and lung that have similar radii $R_{\text{data}} = 1.78 \pm 0.02$ and 1.77 ± 0.02, respectively (Figure 10B-C). Finally, the smallest radius is observed for mouse embryonic stem cells with $R_{\text{data}} = 1.15 \pm 0.04$ (Figure 10E). Because hyperbolic radius indicates the depth of the underlying hierarchical tree, these findings indicate an interesting progression in complexity with embryonic cells exhibiting the smallest degree of hierarchical organization and brain cells exhibiting the largest degree.

We note that the HMDS methods produce estimates of the hyperbolic radius that depend on the embedding dimension D. This happens because the density of points increases exponentially with exponent $(D-1)R$ according to Equation (S3). Results in Figure 10A-E are obtained for a 5D hyperbolic space. In panel F, we show how the estimates of R_{data} decrease with embedding dimension in different data sets (Figure 10F). Importantly, the relative differences in R_{data} across cell types are maintained across a range of different embedding dimensions. The hyperbolic maps continue to have the smallest radius for mouse embryonic cells, larger values for mouse differentiated cells, and yet larger values for mouse brain and human cells. We also find that minimal embedding dimension for all of these datasets is $D = 3$, and smaller dimension fails to properly embed the data.

We also tested robustness of the HMDS method to noise in the data. Towards that goal, we add varying amounts of the multiplicative Gaussian noise to the Lukk et al. data [62] and fit the resulting Shepard diagrams. The fits produce stable convexity estimates for Shepard's diagrams over a broad range of noise values (Figure S5A-E). This robustness is observed up to very large noise values with $\varepsilon = 0.5$ when noise completely destroys the data structure(Figure S5F). The reason for this robustness is that noise does not systematically shift the shape of the Shepard diagram, yielding the same fitting exponent under varying noise amounts.

2.3.5 Hyperbolic low dimensional visualization of gene expression data

While MDS embedding can be used to detect intrinsic geometry, it is not ideal for low dimensional visualization. One of the primary reasons that is common to all MDS-based algorithms is that they are not designed to attract similar points together like t-SNE. Consequently, MDS-based methods achieve poor clustering results. These limitations were solved by non-linear methods like t-SNE and UMAP, which however, are only performed in the Euclidean space. As a result, existing visualization methods may cause distortion of global structure in the data that has a global hyperbolic structure. Here we aim to adapt the t-SNE algorithm to work in hyperbolic

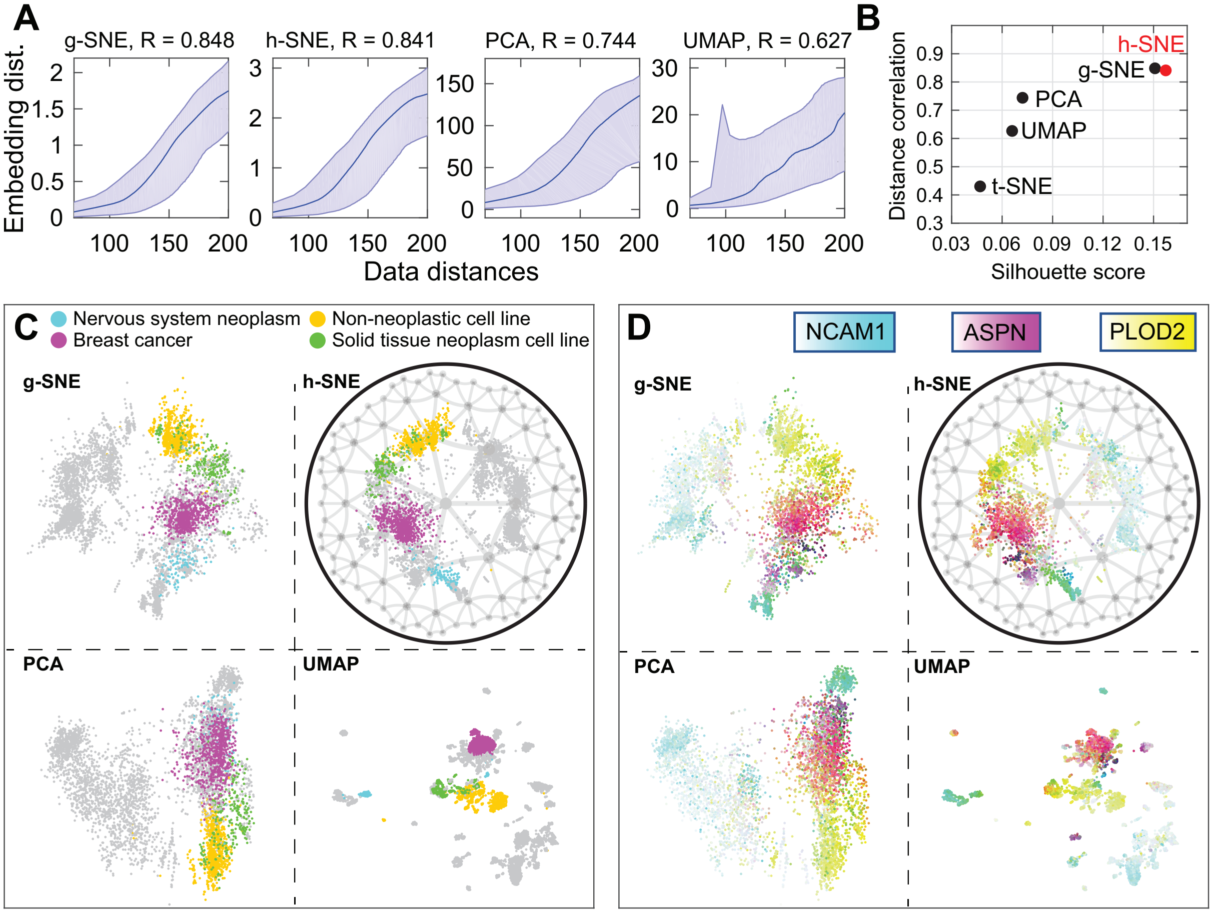

Figure 11. Comparison of two-dimensional visualizations of human expression data [62] using g-SNE, h-SNE, PCA and UMAP. (A) Shepard diagrams of the four different mappings. The Pearson correlation coefficients of pairwise distances plots are shown at the top of each panel. (B) Quantification of local and global structure preservation using Pearson correlation coefficient of Shepard diagram (y-axis) and Silhouette score (x-axis), respectively for five algorithms (including t-SNE). The Silhouette score is defined as the geometric mean of the three scores obtained by using six hematopoietic labels, four malignancy labels and fifteen subtype labels respectively (see Methods). (C) In h-SNE, the data points are visualized within a 2D Poincaré disk with order-7 triangular tiling, which represents a compressed version of a hyperbolic space. In g-SNE, PCA and UMAP, the points are visualized in 2D Euclidean plane. Four cell types are highlighted with color: nervous system neoplasm (cyan), breast cancer (magenta), solid tissue neoplasm cell line (green) and non-neoplastic cell line (yellow). The rest of data points are shown in gray to avoid confusion between multiple colors, see Figure S6 for colors across all cell types. (D) The embedding samples are colored using subtractive CMY color mode according to normalized expressions of three marker genes NCAM1 (nervous system neoplasm), ASPN (breast cancer) and PLOD2 (non-neoplastic cell line). See also Figure S6.

space. To achieve this we use hyperbolic metric to evaluate global distances in the data while keeping the local clustering aspects of the algorithm. The standard t-SNE method effectively discards large distance information between distant points. We recently proposed a variant of t-SNE which aims to preserve global Euclidean structure in the data, which was called global t-SNE (g-SNE) [58]. g-SNE method works by adding to the similarity distance measures present in the t-SNE another term that focuses on large Euclidean distances (see Methods). When applied to Lukk et al. data [62], g-SNE preserves data distances very well (Figure 11A, $R = 0.848$). Despite the high quality of embedding, g-SNE cannot reveal the hierarchical structure of data which is only visible in hyperbolic embedding. Therefore, considering that human gene expression space is locally Euclidean and globally hyperbolic, we develop a hyperbolic t-SNE (h-SNE) method that applies hyperbolic metric to global similarities as defined in g-SNE [58], while still using Euclidean metric for original local similarities. We find that h-SNE gives similar embedding accuracy as g-SNE, both of which largely outperform PCA and UMAP, with $R = 0.841$ for h-SNE compared to $R = 0.744$ for PCA and $R = 0.627$ for UMAP (Figure 11A). The distance correlation of Shepard diagram generally quantifies the quality of embedding with respect to large distances, i.e. the global inter-class structure preservation. To measure the local structure preservation, we use the silhouette score which measures the quality of clustering [70]. Here we find that h-SNE achieves higher silhouette score than g-SNE and significantly higher score than other algorithms(Figure 11B).

These quantitative improvements by h-SNE are also reflected in the improved local and global visualizations that the method provides. For local visualization, the clusters identified by h-SNE are well separated with respect to 15 different tissues and disease types (Figure S6). By comparison, the PCA representation does not separate the fifteen clusters very well, mixing nervous system neoplasm cells (cyan) with the breast cancer cells (magenta)(Figure 11C). The non-neoplastic cell line (yellow) are also not separated in the PCA representation from the solid tissue neoplasm cell line (green) (Figure 11C, see all the 15 labels in Figure S6). The UMAP methods separate clusters better but generate too many disconnected components that are difficult to be matched to sample labels (Figure S6). In terms of global properties, the h-SNE visualization generates a clearer global hierarchical organization of clusters which is not attainable in g-SNE embedding: cells from nervous system neoplasm, breast cancer, non-neoplastic cell line and solid tissue neoplasm cell line are sequentially positioned at different branches in the disk (Figure 11C); in addition, the two principal hematopoietic and malignancy axes can be clearly identified in h-SNE, but not in UMAP (Figure S7). Finally, it is particularly interesting to note the differences in hierarchical positioning that are assigned to breast cancer cells (magenta). Many of these cells occupy points with smaller radii. Positions that are closer to the center of the hyperbolic space typically correspond to more de-differentiated cells, as we have already seen in the comparison between mouse embryonic cells and differentiated cells. Thus, the more central positions assigned to breast cancer cells are consistent with observations of them being close to de-differentiated cells.

The quality of h-SNE visualization is also illustrated by the topography with respect to gradient expressions of three marker genes: NCAM1 [71] for nervous system neoplasm, ASPN [72] for breast cancer, and PLOD2 [73] for non-neoplastic cell line. These marker genes are highly expressed in distinct but continuous branches in h-SNE; by comparison, the expression patterns of these three genes are more difficult to organize in g-SNE, to cluster in UMAP, or to separate in PCA (Figure 11D).

In addition to visualizing discrete data, hyperbolic embedding is especially useful in representing temporally continuous data and predicting lineage information. Klimovskaia, et al. [41] developed Poincaré map method to visualize hierarchies in single-cell data. This method used similar idea as t-SNE but implemented hyperbolic metric in the representation space. This has lead to improvements in the representations of cell trajectories. However, the Poincaré map method, being based on t-SNE, still largely discards large distance information. This problem can be well solved by h-SNE which is designed to capture global hyperbolic structure. For comparison with the Poincaré map method(Figure 4 in [41]), we select the mouse hematopoiesis data in Moignard et al. [74]. This dataset consists of cells from different development stages:

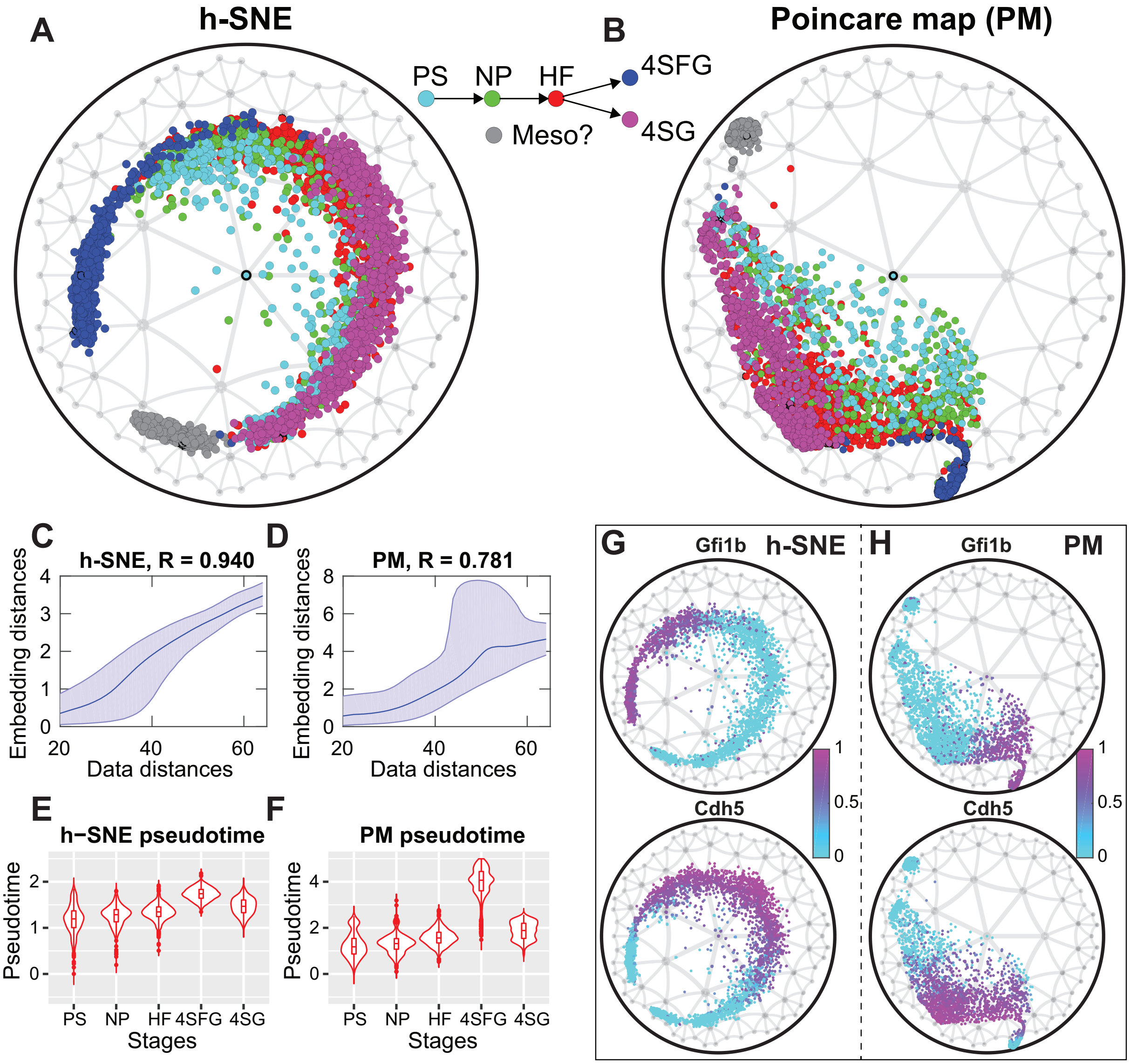

Figure 12. Comparison of hyperbolic embedding of mouse hematopoiesis data using h-SNE and Poincaré map. (A-B) h-SNE and Poincaré mapping of the data after centering the root node in Poincaré disk(see Methods). Gray cluster represents potential outliers or "mesodermal" cells [41][74] . (C-D) Shepard diagram of h-SNE and Poincaré mapping. The data distances are calculated using Euclidean metric in the original data space, while the embedding distances are calculated using hyperbolic metric. (E-F) Predicted pseudo time from h-SNE and Poincaré mapping, the pseudo time is defined as the hyperbolic distances between points and the root node. (G-H) Normalized gene expressions of two main genes Gfi1b (hemogenic marker) and Cdh5 (endothelial marker) in h-SNE and Poincaré maps.

primitive streak (PS), neural plate (NP), head fold (HF), four somite GFP (Runx1) negative (4SG-) and four somite GFP positive (4SG+). We first apply HMDS method to determine the intrinsic geometry of the data and find that the data space is hyperbolic with $R_{\text{data}} = 1.72$ (Figure S8). Then we apply h-SNE to the data and compare the results with Poincaré map. The h-SNE method produces similar local clustering as in Poincaré map, but it generates very distinct global pattern: the two differentiated branches 4SFG and 4SG extend around the disk with clear division along the angular variable in the h-SNE visualization (Figure 12A). The corresponding pattern is not as clear in Poincaré map (Figure 12B). The Shepard diagrams of the embeddings show that h-SNE preserves data distances much better than Poincaré map, especially the large distances (Figure 12C-D). When predicting the pseudo time, the h-SNE method produces a clear pseudo time prediction with a much smaller variance compared to the Poincaré map (c.f. compare the pseudo time in 4SFG stage, Figure 12E-F). Finally, as another example, we show the normalized gene expressions of two marker genes Gfi1b (hemogenic marker) and Cdh5 (endothelial marker), finding that these two genes are differentially expressed in different branches in h-SNE (Figure 12G). This separation is not obvious in Poincaré map (Figure 12H). The clear hierarchical organization of cells in h-SNE map may help us better understand the relationships between cells at different stages.

2.4 Discussion

In this paper we developed a non-metric MDS in hyperbolic space, and showed how it can be used to detect the hidden geometry of data starting with an initial Euclidean representation. By applying this method to several gene expression datasets, we found that gene expression data exhibits Euclidean geometry locally and hyperbolic geometry globally. The radius of the hyperbolic space differed depending on the cell types. The lowest values were observed for embryonic cells and the highest values observed for brain cells in mouse data. Given that hyperbolic geometry is indicative of hierarchically organized data [40, 7], and the spanned radius represents the depth of the network hierarchy, it is perhaps intuitive that the largest value would be observed for highly differentiated and specialized brain cells and the smallest value for the embryonic cells.

The method that we used to detect the presence of hyperbolic geometry was based on non-metric MDS. One can also use methods from algebraic topology [40, 11] for this purpose, as has been recently demonstrated for metabolic networks underlying natural odor mixtures produced by plants and animals [40]. The advantage of the topological method is that it is very sensitive to changes in the underlying geometry, including its dimensionality and hyperbolic radius. However, this method is computationally intensive and does not scale well to large datasets. In contrast, the non-metric MDS method is computationally much faster. Therefore, we recommend to use it as a first step in determining whether the underlying geometry is hyperbolic or Euclidean. If hyperbolic geometry is detected, then radial position of embedding points can be used to arrange data hierarchically. We have also seen that taking into account hyperbolic geometry produces better low-dimensional visualizations, cf. Figure 11 and Figure 12 .

Accurate representation of data across scales is a very active area of research [75, 76, 77]. Special attention is being devoted to developing visualization methods that can not only cluster data in a useful way but also preserve relative positions between clusters [58, 59]. In particular, preserving global data structure was one of the driving factors for the UMAP method [59]. Knowing the underlying geometry helps to position clusters appropriately and robustly map them across different runs in a visualization method. For example, the t-SNE method produces random positions of the clusters across different runs of the algorithm [78]. This problem can in part be alleviated by additional constraints on large distances [58]. Here we find that using a combination of a hyperbolic metric for large distances and Euclidean metric for local distances offers strong improvements in this respect. It also outperforms the recent Poincaré map method that implements hyperbolic metric only for local distances [41]. We notice that although h-SNE is best fit for hyperbolic data, it performs similarly as g-SNE in accuracy distances preservation. It's future direction to further optimize h-SNE algorithm.

What could be the origin of hyperbolic geometry at the large scale and Euclidean at small scale? First, any curved geometry, including hyperbolic, is locally flat, i.e. Euclidean. The scale at which non-Euclidean effects become important depends on the curvature of the space. From a biological perspective, the Euclidean aspects can arise from intrinsic noise in gene expression [63, 64, 65]. This noise effectively smoothes the underlying hierarchical process that generates the data. We find that hyperbolic effects of human gene expression can be detected by including measurements on as few as ~ 100 probes. Why do hyperbolic effects require measurements along multiple dimensions? The reason is that hyperbolic geometry is a representation of an underlying hierarchical process, which generates correlations between variables. These correlations become detectable above the noise once a sufficient number of measurements is made. As an example, one can think of leaves in a tree-like network, and how their activity becomes correlated when it is induced by turning on and off branches of the network. Intuitively, these correlations generate the outstanding branches of a hyperbola. We observe that these correlations can be detected by monitoring even a relatively small (~ 100) number of probes. This makes it possible to construct a global map of genes from partial measurements, and open new ways for combining data from different experiments.

2.5 Methods and Materials

Metric and non-metric MDS in Euclidean model

Assume there sectionre *n* objects described by a set of measurements, the dissimilarity of the objects can be obtained by the experimental measurements of the objects. For example, the dissimilarities of two cells can be calculated by the Euclidean distances of the gene expression vectors. Metric MDS approximates the geometric distances d_{ij} to the data dissimilarities δ_{ij}, while non-metric MDS approximates a monotonic transformation of dissimilarities of data. The transformed values are known as disparities $\hat{d_{ij}}$. The loss function S in Euclidean embedding was defined as :

$$S = \sqrt{\frac{S^*}{T^*}} \tag{S1}$$

Where $S^* = \Sigma_{i,j}(d_{ij} - \hat{d_{ij}})^2$, $T^* = \Sigma_{i,j}d_{ij}^2$. In non-metric MDS, $\hat{d_{ij}}$ is determined using the greatest convex minorant method in Kruskal's approach [79]. In metric MDS, disparities are equal to dissimilarities: $\hat{d_{ij}} = \delta_{ij}$.

2.5.1 Non-metric MDS in native hyperbolic model

There are many hyperbolic space representations, we will use the native representation with polar coordinates [7] in our hyperbolic MDS. The angular coordinates in the space are the same as in an Euclidean ball, the radius $R_{\text{model}} \in (0, \infty)$ characterizes the hierarchical depth of the structure, measures the degree of hierarchy in data, and determines how points distribute in the space. The distance of two points d_{ij} is calculated as:

$$\cosh(d_{ij}) = \cosh(r_i)\cosh(r_j) - \sinh(r_i)\sinh(r_j)\cos(\Delta\theta_{ij}) \tag{S2}$$

Where r_i and r_j are the radial coordinates of the two points, and $\Delta\theta_{ij}$ is the angle between them. In D-dimensional HMDS, we initialize the embedding process by uniformly sampling points within radius R_{model} in the native hyperbolic model. The points directions are uniformly sampled around the high-dimensional sphere, and the radial coordinate $r \in (0, R_{\text{model}}]$ follows :

$$\rho(r) \sim \sinh^{D-1} r \tag{S3}$$

We note that there can be merits to sample the points uniformly in the angular variables. Although this does not lead to uniform sampling of points along the sphere, this way of sampling can be particularly advantageous in the situation where the angular variable maps onto periodic variables that correspond to cell cycle or other rhytms. We have used this sampling in our previous publication on olfactory signals produced by fruits and plants [40] where it matched developmental processes in the fruit.

During the iteration process, we update both angular and radial coordinates according to the gradient descent of the loss function Equation (S1), and at the same time set R_{model} as the upper bound of the radial coordinates. The reason of setting a bound is that the coordinates in hyperbolic model are polar coordinates which cannot be normalized after each iteration as performed in Euclidean MDS, so without bound the gradient descent of loss functionsection Equation (S1) would lead to very large r_i and d_{ij} (since d_{ij} is in the denominator) and hence fail to preserve radial coordinates of data. By setting the upper bound for radial coordinates, the HMDS embedding can well preserve the data distances and precisely detect hyperbolic radius of data R_{data} (Figure S3).

2.5.2 Fitting of Shepard diagram

The Shepard diagram is linear if the geometry of input data matches the geometry of embedding space, and otherwise nonlinear. In both EMDS and HMDS, we use the power function below to fit the pairwise distances:

$$y = a(x - x_0)^{\kappa+1} \tag{S4}$$

Where $x_0 = \min(x) - \varepsilon$ is an offset representing the distance caused by intrinsic noise of data, a small value ε is introduced to avoid zero inputs in the fitting. The convexity κ describes the linearity of the fitting. $\kappa = 0$ indicates Euclidean input in EMDS and means $R_{\text{data}} = R_{\text{model}}$ in HMDS. $\kappa > 0$ means the data is more hyperbolic than the model, and vice versa.

2.5.3 Hyperbolic t-SNE

Given a data set containing N data points described by D dimensional vectors $\{\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3, ..., \mathbf{x}_N; \mathbf{x}_i \in \mathbb{R}^D\}$, the t-SNE algorithm [57] defines the similarity of two points $\mathbf{x}_i, \mathbf{x}_j$

as the joint probability p_{ij}:

$$p_{j|i} = \frac{\exp(-\|\mathbf{x}_i - \mathbf{x}_j\|^2/2\sigma_i^2)}{\sum_{k\neq i}\exp(-\|\mathbf{x}_i - \mathbf{x}_k\|^2/2\sigma_i^2)}, \tag{S5}$$

$$p_{ij} = \frac{p_{j|i} + p_{i|j}}{2n}. \tag{S6}$$

The similarity of two points $\mathbf{y}_i, \mathbf{y}_j$ in embedding space is defined as the joint probability q_{ij}:

$$q_{ij} = \frac{(1+\|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{m\neq n}(1+\|\mathbf{y}_m - \mathbf{y}_n\|^2)^{-1}}. \tag{S7}$$

The discrepancy between the similarities of data and embedding points is the loss function, which is defined by Kullback-Leibler (KL) divergence of the joint probability p_{ij} and q_{ij} :

$$L = D_{KL}(P\|Q) = \sum_i \sum_j p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right). \tag{S8}$$

Minimizing the loss function L with respect to the embedding coordinates $\mathbf{y}_i$ by gradient descent gives:

$$\frac{\partial L}{\partial \mathbf{y}_i} = 4\sum_j (p_{ij} - q_{ij})(\mathbf{y}_i - \mathbf{y}_j)(1+\|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}. \tag{S9}$$

The original definitions of similarities Eqs. (S5-S7) in t-SNE are sensitive to small pairwise distances among neighboring points but not to large distances between distant points. To preserve large distances, Zhou et al. [58] proposed global t-SNE algorithm that introduced global similarity terms $\hat{p}_{ij}$ and $\hat{q}_{ij}$ which are primarily sensitive to large distance values:

$$\begin{aligned}\hat{p}_{ij} &= \frac{1+\|\mathbf{x}_i-\mathbf{x}_j\|^2}{\sum_{m\neq n}(1+\|\mathbf{x}_m-\mathbf{x}_n\|^2)}\\ \hat{q}_{ij} &= \frac{1+\|\mathbf{y}_i-\mathbf{y}_j\|^2}{\sum_{m\neq n}(1+\|\mathbf{y}_m-\mathbf{y}_n\|^2)}\end{aligned} \tag{S10}$$

And they defined the global loss function $\hat{L}$ as:

$$\hat{L} = D_{KL}(\hat{P}\|\hat{Q}) = \sum_i\sum_j \hat{p}_{ij}\log\left(\frac{\hat{p}_{ij}}{\hat{q}_{ij}}\right) \tag{S11}$$

The total loss function L_{total} was then defined by combining the two loss functions using a weight parameter λ:

$$L_{\text{total}} = L + \lambda\hat{L} \tag{S12}$$

The gradient of the total loss function L_{total} gives:

$$\frac{\partial L_{\text{total}}}{\partial \mathbf{y}_i} = 4\sum_j[(p_{ij}-q_{ij})-\lambda(\hat{p}_{ij}-\hat{q}_{ij})]\cdot(\mathbf{y}_i-\mathbf{y}_j)(1+\|\mathbf{y}_i-\mathbf{y}_j\|^2)^{-1}, \tag{S13}$$

where the weight λ of the global loss function controls the balance between the local clustering and global organization of the data. Large λ values lead to more robust global distribution of clusters, but less clear classifications. Small λ moves back to approximate the traditional t-SNE, and will be exactly the same when $\lambda = 0$. In hyperbolic t-SNE, we still use native representation parametrized by R_{model} as in HMDS, but here R_{model} is only used to determine the initial radial distribution, not to set the upper bound. We substitute the Euclidean distances in global similarity terms Equation (S10) by hyperbolic distances d^h_{ij} defined in Equation (S2), and change Cartesian coordinate system to polar one for all the distance calculations. Then the gradient of total loss function with respect to polar coordinates would be:

$$
\begin{aligned}
\frac{\partial L_{total}}{\partial r_i} &= 4\sum_j [(p_{ij}-q_{ij})\cdot d_{ij}^e \cdot \frac{\partial d_{ij}^e}{\partial r_i}(1+(d_{ij}^e)^2)^{-1} \\
&\quad -\lambda(\hat{p}_{ij}-\hat{q}_{ij})\cdot d_{ij}^h \cdot \frac{\partial d_{ij}^h}{\partial r_i}(1+(d_{ij}^h)^2)^{-1}] \\
\frac{\partial L_{total}}{\partial \theta_i} &= 4\sum_j [(p_{ij}-q_{ij})\cdot d_{ij}^e \cdot \frac{\partial d_{ij}^e}{\partial \theta_i}(1+(d_{ij}^e)^2)^{-1} \\
&\quad -\lambda(\hat{p}_{ij}-\hat{q}_{ij})\cdot d_{ij}^h \cdot \frac{\partial d_{ij}^h}{\partial \theta_i}(1+(d_{ij}^h)^2)^{-1}]
\end{aligned} \tag{S14}
$$

Where d_{ij}^e is the Euclidean pairwise distance in polar coordinates, d_{ij}^h is the hyperbolic pairwise distance obtained from Equation (S2). p_{ij}, q_{ij}, $\hat{p}_{ij}$ and $\hat{q}_{ij}$ are defined by Eqs. (S5-S7) and Equation (S10) with polar coordinates. When implementing the algorithm, we substitute the radial coordinates with their exponential transformation to avoid negative radii during the iteration:

$$
r_{exp} = e^r \tag{S15}
$$

The derivative of distances with respect to the new variable would be:

$$
\frac{\partial}{\partial r_{exp}} = \frac{\partial}{\partial r}\cdot\frac{\partial r}{\partial r_{exp}} = \frac{1}{r_{exp}}\cdot\frac{\partial}{\partial r} \tag{S16}
$$

When the iterations converge, we make logarithm transformation of r_{exp} to get the real radial coordinates.

2.5.4 Parameters in visualization algorithms

For Lukk et al. [62] data, we set $\lambda = 8$ and $R_{\text{model}} = 1$ in h-SNE. We select the result that best preserves data distances from 30 repeats. After obtaining the embedded points, we transform the points from native representation to Poincaré disk model by performing the transformation on radial coordinates:

$$r_{\text{Poincare}} = \tanh(\frac{r_{\text{native}}}{2}) \tag{S17}$$

In g-SNE, the parameter is: $\lambda = 20$. In PCA, we use the first two principal components for visualization. In UMAP, we screen a wide range of the combination of two key parameters: number of neighbors $\in \{5, 10, 20, 50, 100\}$ and minimal distance $\in \{0.001, 0.01, 0.1, 0.5, 0.8\}$, each of the 25 combinations was repeated 30 times. The optimal combination of parameters that leads to largest distance correlation of Shepard diagram is: number of neighbors $= 100$ and minimal distance $= 0.5$, and the corresponding result is shown in Figure 11. For Moignard et al. data [74], the parameters for h-SNE are: $\lambda = 10$, $R_{\text{model}} = 1$. The root node index is 1800. When plotting Poincaré map, we directly use the embedding positions provided in Klimovskaia et al. [41].

2.5.5 Evaluation of embedding

The Pearson correlation coefficient of Shepard diagram (embedding distances versus data distances) is used to measure the preservation of distances and global structure. For local clusters, we apply silhouette score [70] to our embedding results. Silhouette score measures the quality of data partitioning and clustering in graphical representation of objects, which in our case can be used to measure the consistency of the data configuration in 2D embeddings with the "ground truth" cluster labels. The higher score indicates better consistency with data labels. We consider all the three types of labels available – six hematopoietic properties, four malignancy properties and fifteen subtypes, and calculate the geometric mean of silhouette scores obtained by using these three labels:

$$s = \sqrt[3]{s_1 s_2 s_3} \tag{S18}$$

Where s_1, s_2, s_3 represent the silhouette scores by using the three types of labeling respectively. The mean score s is used to quantify the local structure preservation for the five visualization algorithms, which is shown in Figure 11B.

2.5.6 Data preprocessing and analysis

No pre-processing was done for the microarray dataset from human samples [62].

For scRNA-seq dataset from [67], we use Seurat packages [69] to perform normalization, feature selection and scaling for the data, and to select the top 50 principal components for analyses. These results are reported in Figure 10. The results without pre-processing (and keeping all 1000 principle components) were qualitatively similar but had slightly reduced hyperbolic radii: $R_{\text{mouse brain}} = 1.62 \pm 0.02$, $R_{\text{mouse lung}} = 1.47 \pm 0.03$, $R_{\text{mouse kidney}} = 1.45 \pm 0.03$, and $R_{\text{mouse embryo}} = 0.98 \pm 0.02$ (compare with number in Figure 10). These obervations are consistent with the observation that noise reduction done during pre-processing makes hyperbolic effects stronger and more apparent.

For mouse hematopoiesis data, we use the processed data from Klimovskaia et al. [41]. The Seurat analysis, violin plots and linear regression were performed using R version 3.6.2, the other analyses were performed using MATLAB R2017a.

2.6 SUPPLEMENTARY MATERIALS

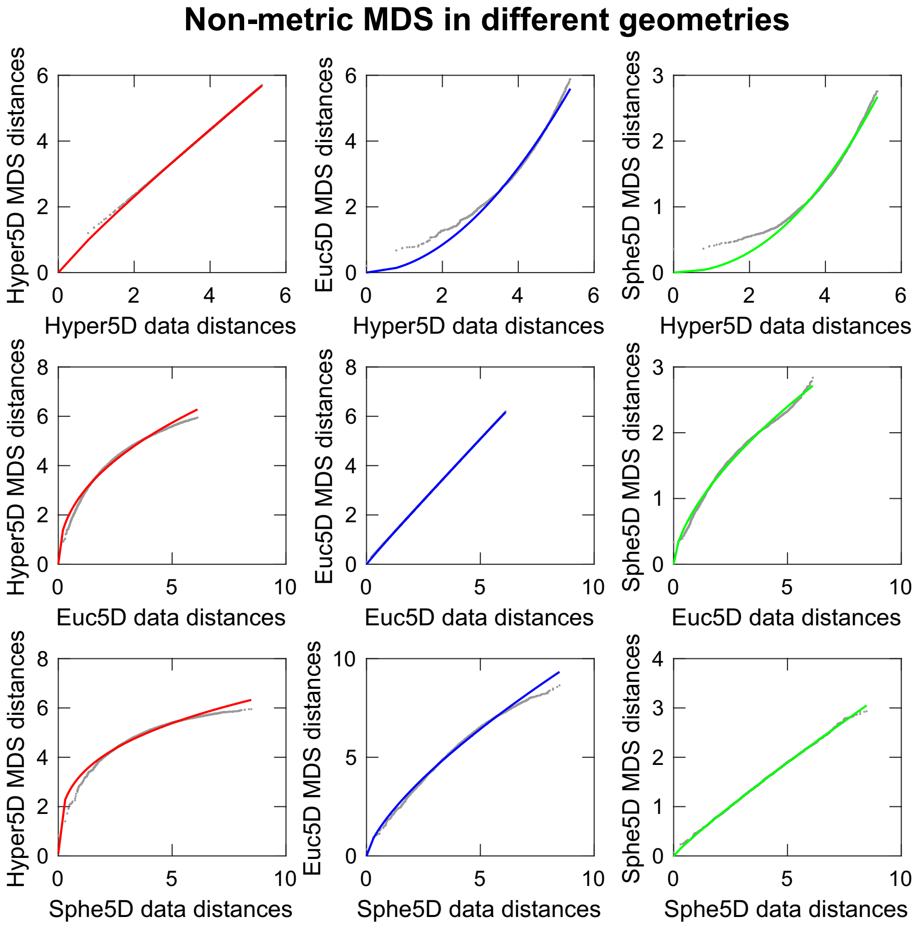

Figure S8. Illustration of non-metric MDS embedding in different geometries. Related to Figure 7. 100 synthetic points in 5D hyperbolic (top), Euclidean (middle) and spherical (bottom) space are embedded into 5D hyperbolic (left), Euclidean (middle) and spherical (right) space respectively. The hyperbolic radius is $R = 5$ for both data and model. The fitting methods are the same as in Figure 2 in the manuscript.

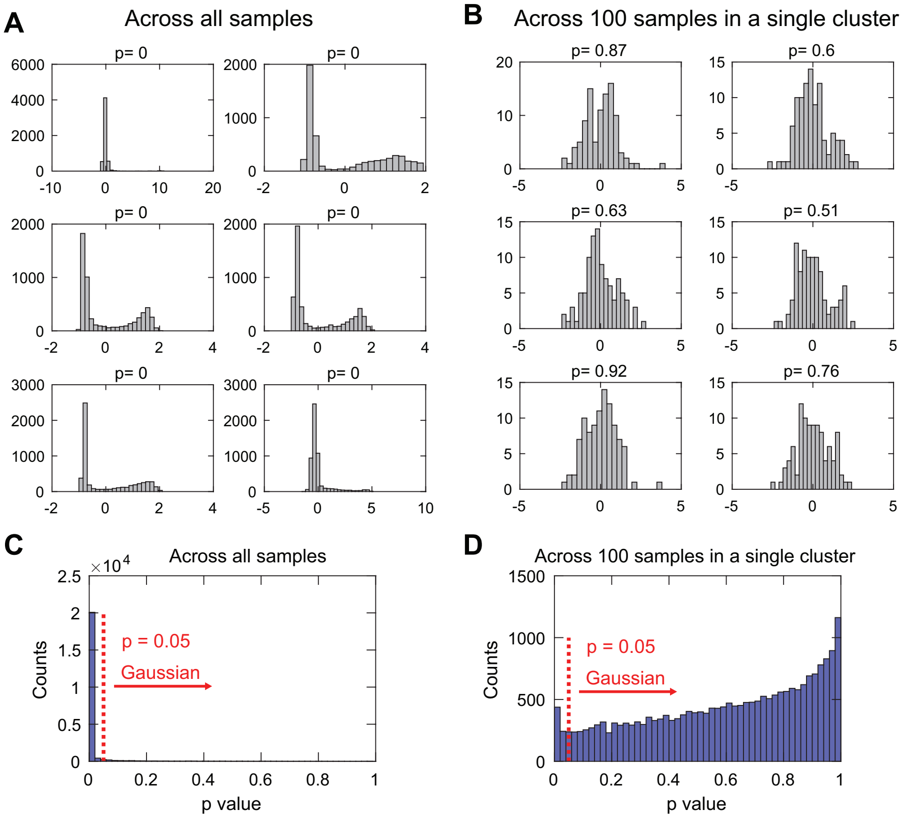

Figure S9. Gaussianity of normalized gene expressions across the whole samples and from a single cluster. Related to Figure 8. (A) Gene expression distributions of the six most non-Gaussian distributed probes across all the samples, p values were given by one-sample Kolmogorov-Smirnov test for Gaussianity, the null hypothesis is that the normalized distribution is standard normal distribution. (B) Gene expression distributions of the same six non-Gaussian probes as in (A) across 100 samples in one of the k-means ($k = 80$) cluster. (C) p value distributions of all the probes across the whole samples. (D) p value distributions of all the probes across 100 samples in one of the k-means cluster.

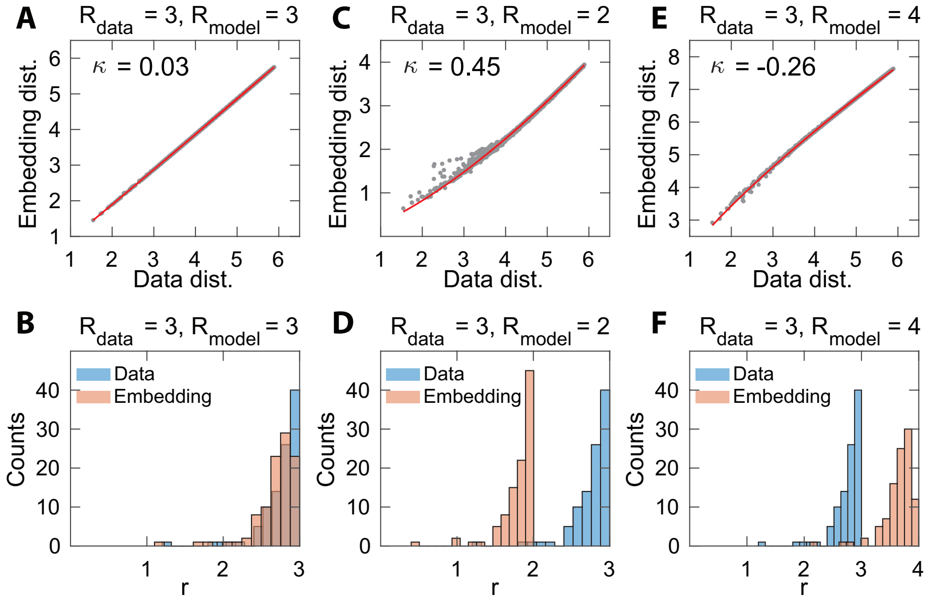

Figure S10. HMDS embedding of hyperbolic data with different R_{model}. Related to Figure 10. 100 points are sampled in hyperbolic space with $D = 5$, $R_{\text{data}} = 3$. The embedding dimension is $D = 5$ in (A-F). The Shepard diagram convexity κ is shown in panel (A,C,E). (A) Shepard diagram of HMDS embedding of the samples to 5D hyperbolic space with $R_{\text{model}} = 3$. (B) Histogram of radial coordinates r of 100 sample points and model points after HMDS embedding with $R_{\text{model}} = 3$. (C-D) $R_{\text{model}} = 2$. (E-F) $R_{\text{model}} = 4$.

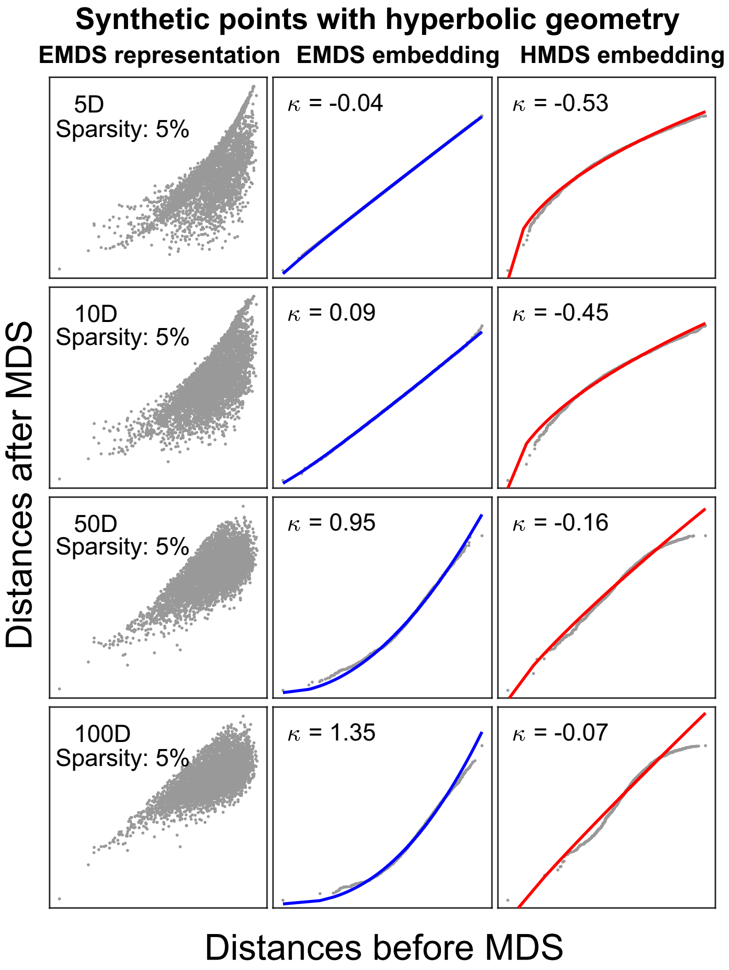

Figure S11. Robustness to changes in data sparseness on geometry detection. Related to Figure 7 and Figure 10. In Euclidean representation after EMDS, the coordinates of all the points are thresholded by fifth percentile of all the coordinate values to simulate the sparse RNA-seq values of cells. The left column shows the plots of thresholded embedding distances versus distances before MDS. The fitting plots and inserted convexity values in the rest columns have same meanings as in Figure 2 in the manuscript.

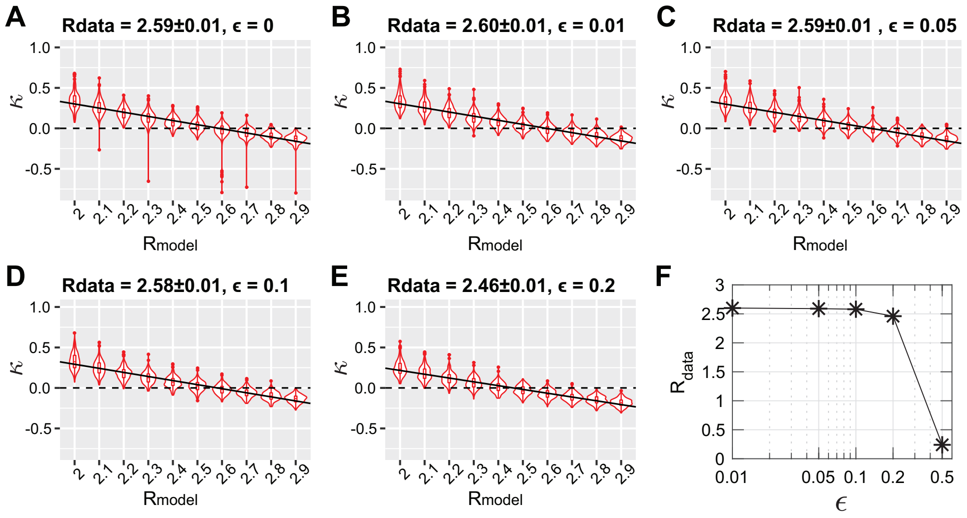

Figure S12. Screening R_{data} of Lukk data with different magnitude of noise added by doing HMDS. Related to Figure 10. The embedding dimension is $D = 5$. The noise ε is added as multiplicative Gaussian noise: $M_{noise} = M[1 + \varepsilon \cdot N(0,1)]$. (A-E) Fitting of R_{data} under different ε. (F) Plot R_{data} as the function of ε.

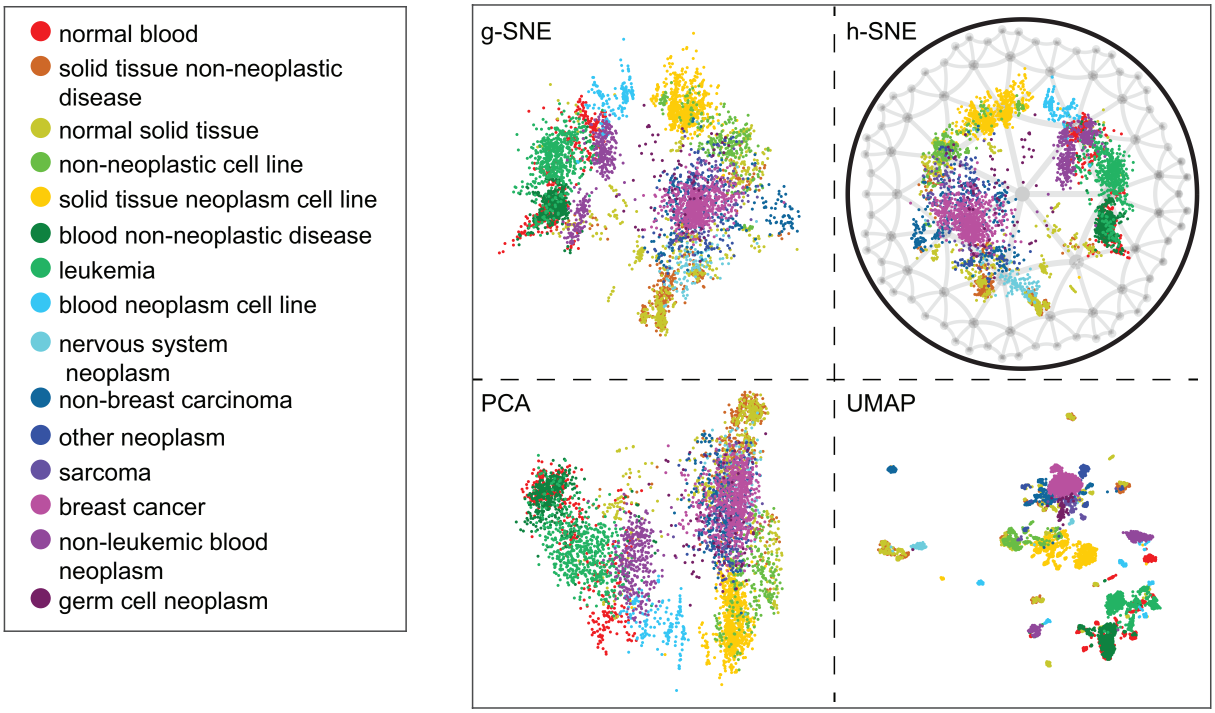

Figure S13. Comparison of two-dimensional visualizations of human expression data using g-SNE, h-SNE, PCA and UMAP. Related to Figure 11. The samples are colored according to the 15 tissue and disease types.

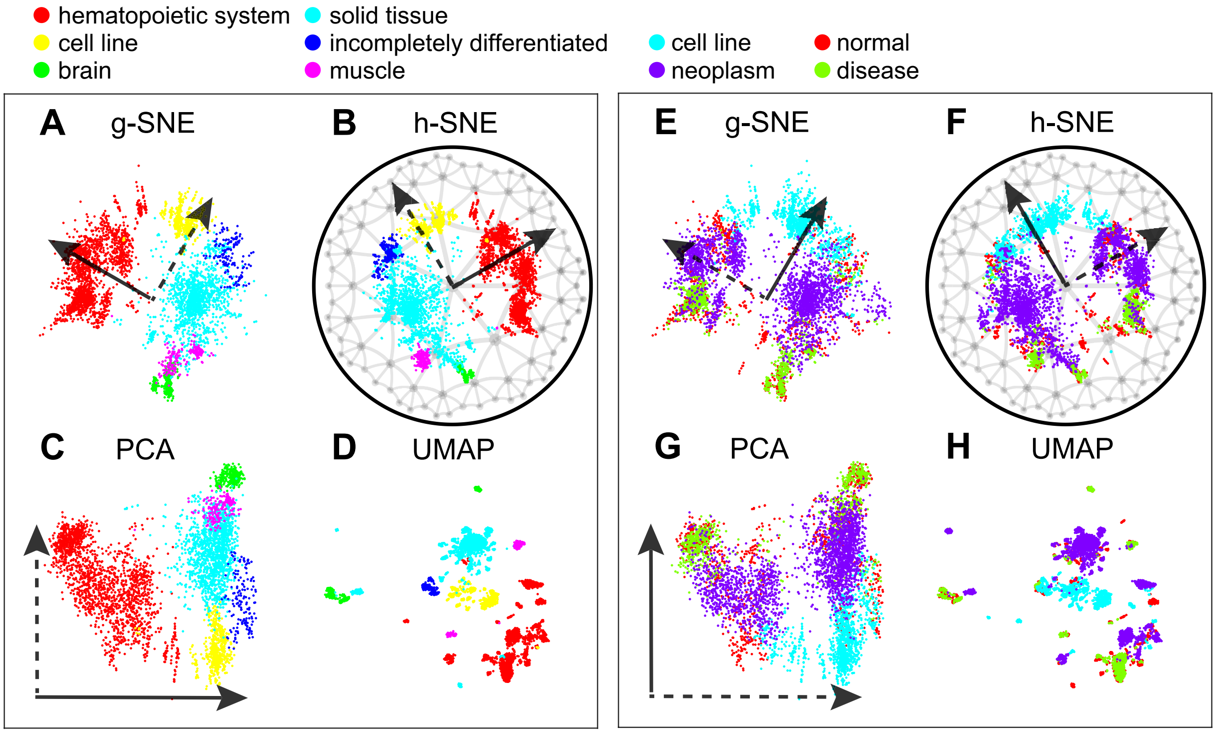

Figure S14. Comparison of two-dimensional visualizations of human gene expression data in different algorithms. Related to Figure 11. (A-D) g-SNE, h-SNE, PCA and UMAP embeddings of human samples classified by hematopoietic properties, these labels also represent the six major clusters identified by Lukk et al. The hematopoietic axes are shown in solid lines in g-SNE(A), h-SNE(B) and PCA(C). (E-H) g-SNE, h-SNE, PCA and UMAP embeddings of human samples classified by malignancy properties. The malignancy axes are shown in solid lines in g-SNE (E), h-SNE(F) and PCA(G). The six major clusters, the hematopoietic axis and the malignancy axis are hard to identify in UMAP.

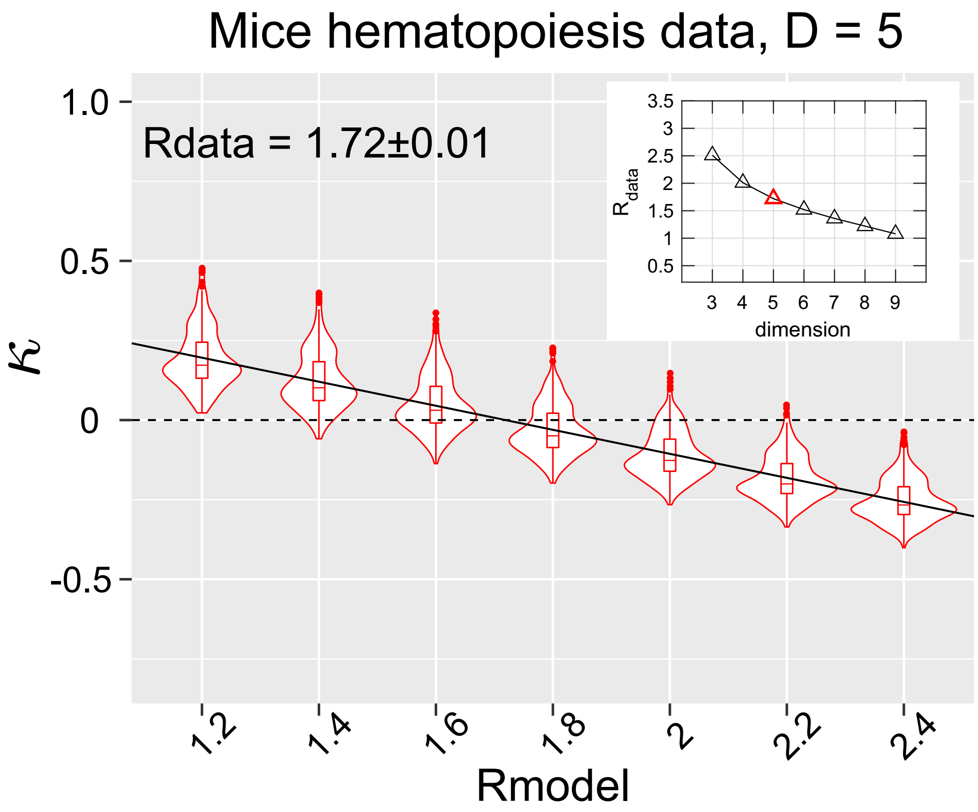

Figure S15. Screening $R_{\mathbf{data}}$ of mice hematopoiesis data in Moignard et al. by doing HMDS. Related to Figure 12. The inset shows the fitted R_{data} as the function of the embedding dimension.

Chapter 2 , in full, is a reprint of the material as it appears in Hyperbolic geometry of gene expression. *Iscience*, 2021; Yuansheng Zhou, Tatyana O Sharpee. The dissertation author was the primary investigator and author of this paper.

Chapter 3

Hyperbolic geometry – a representation inspiring novel insights

The geometric detection and visualization of data has already been introduced in the previous two chapters. In Chapter 1, we first discovered the parameters of hyperbolic geometry of olfactory space, and then used hyperbolic MDS to embed the data points to the corresponding geometry. On one hand, the biological relatedness of the principal axes in the space further validated the visualization; one the other hand, the identified geometric space provided a reliable geometric description of the abstract odorants, which provides a novel approach to organizing odors as well as characterizing the objects containing the odors, such as the fruits. In Chapter 2, we used HMDS to show that the gene expression space of different biological systems has hyperbolic geometry on a global scale, and visualized the human samples in Lukk data using hyperbolic t-SNE. Hyperbolic t-SNE is different from hyperbolic MDS in that it aims to cluster the cells and separate different types, which is just what the single cell RNAseq analysis desires; and the analysis on Lukk data can be easily extended to other types of single cell transcriptome data. In addition to geometry discovery and data visualization, the hyperbolic MDS and hyperbolic t-SNE methods may have broader application in uncovering biological insights. In this chapter, I will discuss the application of the methods to several other datasets and show how hyperbolic geometry inspires new discoveries in different biological systems.

3.1 Hyperbolic stimuli space for olfactory receptor responses

Individual olfactory receptors (OR) respond to many odor ligands, and each odor ligand evokes responses from many ORs. How the activities of ORs collectively encode natural odor mixtures remains an open question. In Chapter 1 we have demonstrated that odor molecules can be mapped onto a three dimensional hyperbolic space based on the statistics of their co-occurrence within natural mixtures, and that the principal perceptual properties of odorants, e.g. pleasantness, can be well represented by the axes in the space. This indicated that the hyperbolic embedding space of natural odorants may serve as the stimuli space for olfactory receptors. To show this we use the concentration measurements of odorants from strawberry and tomato datasets used in Zhou et al.[40], and the OR response datasets from Hallem et al.[80]. We combined strawberry and tomato odor datasets based on their overlapping odorants, and then selected the common odorants that are available in both natural odor datasets and receptor response datasets. The similarities of odorants in terms of OR responses were defined as the Euclidean distances of available receptor activities vectors. The co-occurrence similarities were defined as the absolute values of correlation coefficients of odorant concentrations across samples. The geometric distances of odorants were calculated using both hyperbolic and Euclidean representations of natural odorants, which are achieved by hyperbolic multi-dimensional scaling used in [40] and Euclidean multi-dimensional scaling respectively. Figure 13 shows the correlations between the OR response similarities and odorants stimuli similarities. The correlation is significant when using co-occurrence statistics in natural fruit samples as the stimuli, compared with the shuffling results (Fig. 13A-B) . In the geometric representations, stimuli similarities are given by the geometric distances of the embedding points. Hyperbolic representation leads to a much higher increase of correlation compared with Euclidean representation (Fig. 13C-D). These findings show that OR responses capture the co-occurrence statistics of natural odorants, and that a hyperbolic model is a proper representation of odorants stimuli space for OR responses.

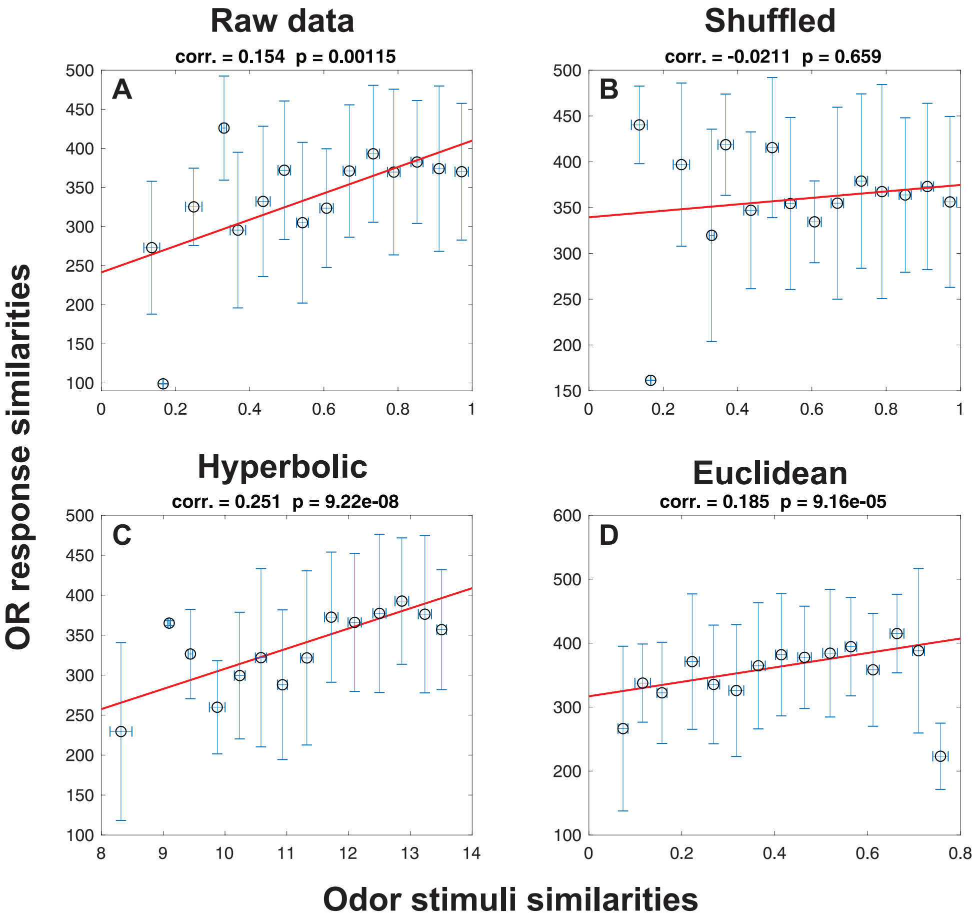

Figure 13. Correlation between pairwise distances of odorants calculated from OR response similarities and from odor stimuli similarities. (A) OR response similarities vs. odor co-occurrence statistics in combined natural odor dataset. (B) OR response similarities after shuffling the odorants vs. odor co-occurrence statistics. (C) OR response similarities vs. geometric distances in hyperbolic representation. (D) OR response similarities vs. geometric distances in Euclidean representation.

3.2 Characterization of structure-specific cell types across different brain regions.

We have shown in Chapter 2 that h-SNE outperforms other algorithms qualitatively and quantitatively in 2D visualization for Lukk data. The Lukk data is a relatively small dataset and has a limited degree of complexity. In this section, we apply h-SNE to a highly complicated dataset which contains scRNAseq measurements of very large number of cells in nine mouse brain regions. The dataset came from Saunders et al [81] in which they used Drop-seq to profile RNA expressions in 69000 cells from nine regions in mouse brain. We analyzed the data in both a global scale which considers cells from the whole brain, and a local scale which focuses on specific brain regions. We first perform t-SNE, UMAP and h-SNE to 4500 cells equally sampled from the nine regions (500 cells in each region) and visualize the global mouse brain atlas in 2D map (Fig.14A-C). Some of the brain regions, such as the striatum and hippocampus, are well separated from other regions in all the three algorithms; while some other regions, such as cerebellum and substantia nigra, are broken into disconnected sub-clusters in t-SNE and UMAP embedding, and only continuously represented in h-SNE (Fig.14A-C, and the small disks in Fig.14D). The global structures of the disconnected components in t-SNE and UAMP are hard to detect, but a branching structure is clearly shown in h-SNE map. In h-SNE map, there are two types of regions which show distinct spatial organizations in the disk. One is called "centering region": the cells from entopedencular(ENP), cerebellum(CB), globus pallidus(GP), substantia nigra(SN) and thalamus(TH) locate around the center of disks; the other type is called "branching region": cells from the frontal cortex(FC), posterior cortex(PC), hippocampus(HC) and striatum(ST) stretch out from the center like branches (Fig.14C). Next we look at each of the nine regions separately and perform h-SNE embedding for 5000 cells sampled from each region. In these regional embeddings, cells are further separated into several sub-clusters [81] and labeled by different colors. The cell distributions in regional mapping are consistent with the distribution in global mapping: in the five "centering regions", different sub-clusters expand in all directions from center in the disks; while in the four “branching regions”, the sub-clusters tend to branch out in a single direction (Fig.14D). Next we study the relationship between the cell structures and cell distributions in the disk. The cells with low gene expressions (few Unique Molecular Identifiers (UMIs)) tend to locate in the center of the disk (CB and TH); the granule cells in CB and dentate gyrus are small neurons and form circular shapes surrounding the centers (CB and HC); cells in CA1 and CA3 of hippocampus and layer 2/3 in frontal and posterior cortex are mostly pyramidal neurons, and they form “crabs” extending to the boundaries of the disks (HC, FC and PC); polydendrocytes are glial cells and distribute like a “narrow path” going from the center to the boundary (ENP and GP). These spatial patterns of different cell types are hard to find in t-SNE and UMAP embedding (Fig.14E). These findings show that structurally similar cell types across brain regions are characterized by similar spatial localization patterns in the 2D hyperbolic disk.

The quality of the embeddings are validated by the quantitative evaluation of data distances preservation. We calculate the correlation coefficient between the pairwise distances of the low dimensional embedding points and original high dimensional data points, and find that h-SNE best preserves the data distances with the highest correlation coefficient across all the nine brain regions in Saunders et al (Fig.15). From the distances plots, we notice that h-SNE not only preserves distances with less noise (narrower shadows), but also preserves the intrinsic geometry of the data, as can be seen in the linear distance relationships in the plots, compared with the other embeddings (Fig.15).

3.3 Characterization of region-dependent cell types across different differentiation stages.

In Section 3.2, we show that h-SNE embedding can characterize structure-specific cell types across brain regions, here we further study whether the method can be applied to characterize region-specific cell types across brain regions in dynamic process, e.g. cell differentiation.

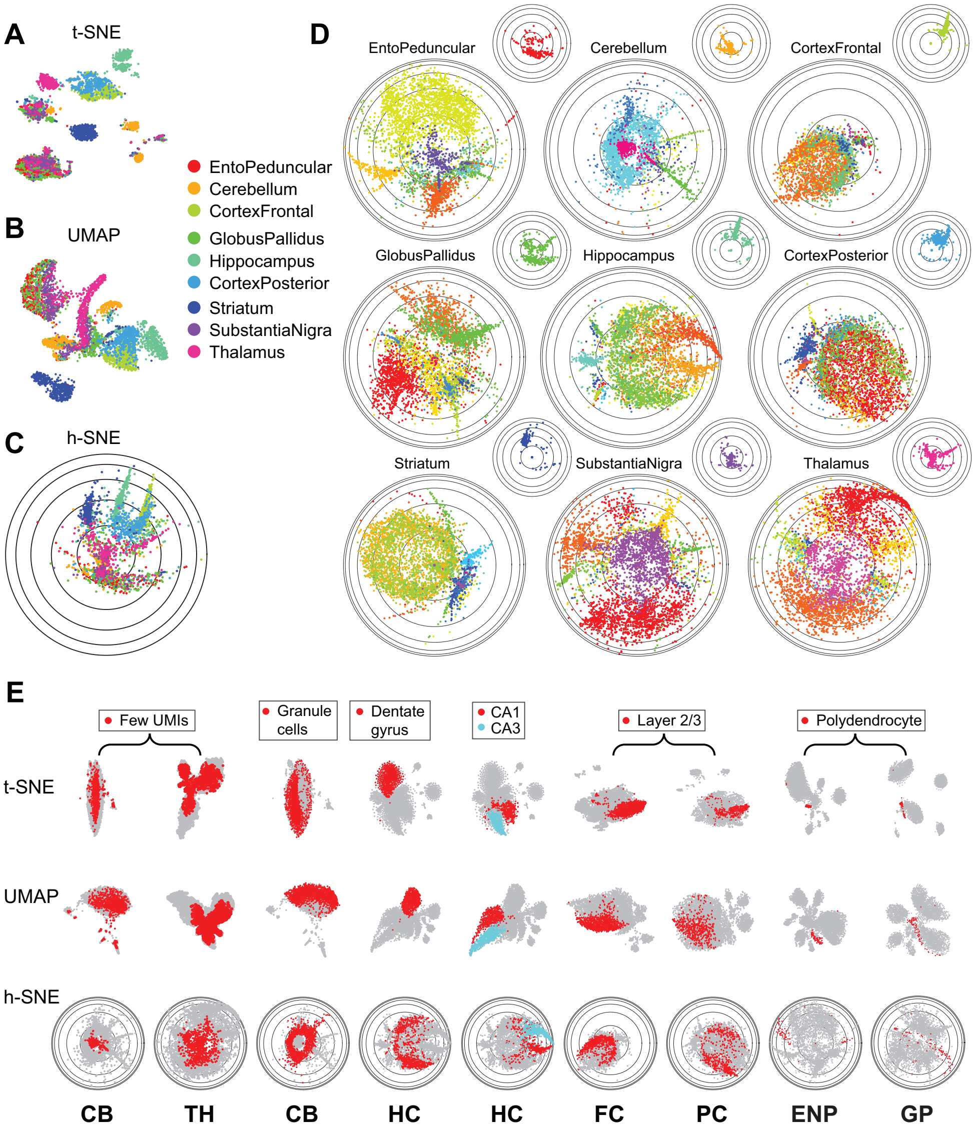

Figure 14. 2D mapping of mouse brain atlas and each of the nine anatomical regions. (A-C) Global map of 4500 cells in mouse brain which includes nine regions (500 cells for each region) using t-SNE, UMAP and h-SNE embedding. The regions are labeled with different colors. (D) The larger disks show the h-SNE mapping of 5000 cells in each of the nine regions, each color represents a cluster annotated in [7]. The smaller disks at the top right corners show the spatial distributions of cells extracted from the global map in (C), the regions are colored the same way as in (C). (E) Specific cell types are marked out using red or light blue colors in t-SNE, UMAP and h-SNE mapping of cells to show their spatial distributions. Few unique molecular identifiers (UMIs) means the cells contain too few UMIs and were not annotated.

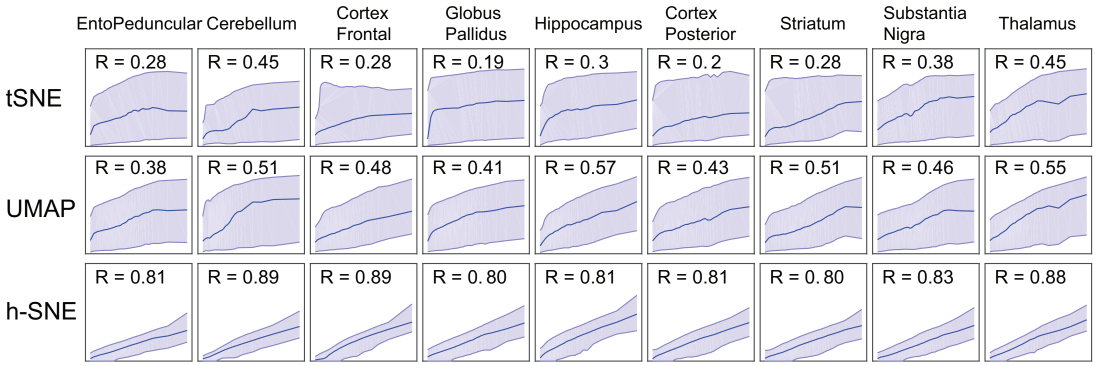

Figure 15. Pairwise distances preservation in embedding for the data from nine regions of mouse brain. The pairwise distances of embedding points are plotted against the data points. The data pairwise distances are grouped using 50 bins. The solid lines and shadows represent the median values and 95% intervals of the embedding distances versus median of data distances in the 50 bins. The Pearson correlation coefficients between embedding distances and data distances are shown in the top of each panel.

Marques[82] performed single cell RNAseq for 5072 cells to study the oligodendrocytes differentiation in 10 regions of mouse central nervous system. They identified 13 cell types including: vascular and leptomeningeal cells (VLMC), oligodendrocyte precursors (OPC), differentiation-committed oligodendrocyte precursors (COPs), two subtypes of newly-formed oligodendrocytes (NFOL1-2), two subtypes of myelin-forming oligodendrocytes (MFOL1-2), six subtypes of mature oligodendrocytes (MOL1-6). These cell types represent different stages of oligodendrocytes differentiation. The authors performed t-SNE and defined the 10 regions to be immature (anterior regions such as amygdala and hippocampus), intermediate (corpus callosum, zona incerta, striatum and hypothalamus) and mature (cortex and posterior regions such as dorsal horn and SN-VTA), based on the mature cells proportions in different regions of juvenile brain. However, the classification of regions was qualitative and the differences of these “mature” regions were not explored. Here we performed t-SNE, UMAP and h-SNE (Fig.16A-C) on the whole dataset (including both juvenile and adult brain), and find that all the three embedding methods show clear global differentiation trajectories, but in totally different ways. t-SNE mapping is similar to the result in [6],where the two sub-clusters MOL1-4 and MOL5-6 are mixed together (Fig.16A). UMAP mapping shows a branch between MFOL and MOL cells, however, this branch does not separate MOL1-4 and MOL5-6 since they exist in both branches(Fig.16B). h-SNE shows a narrow path before MFOL and then “explodes” to circular shape, MOL1-4 locate at the inner circle and MOL 5-6 move further to the outer circle(Fig.16C). Next we separate the cells based on the anatomical regions and study how the mature cells distributions differ in the five regions where mature cells abound: SN-VTA, dorsal horn, hypothalamus, cortex S1 and corpus callosum(Fig.16D-F). t-SNE does not show a clear structure in the mature cells distribution in these five regions (Fig.16D). In UMAP, the mature cells in SN-VTA and corpus callosum locate in different branches, but the cells in the dorsal horn, hypothalamus and cortex S1 distribute in both branches and cannot be distinguished(Fig.16E). In h-SNE, we find that the mature cells of the five regions locate at different radii in the hyperbolic disk(Fig.16F). The median embedding radii of MOL5-6 cells are larger than MOL1-4 in all the five regions, and the median radii of both MOL1-4 and MOL5-6 cells increase from SN-VTA to corpus callosum (Fig.16G). These results show the expression patterns of the same mature cells are region-dependent in an organized way, and may provide new insights of studying cell types in different regions during the differentiation. The quality of the embeddings are also evaluated by distance preservation: the h-SNE embedding performs best with the correlation of $R = 0.78$ in distance plots(Fig.16H).

3.4 Conclusion and future outlook

In this chapter we show some preliminary results to illustrate how the hyperbolic geometry can inspire new insights for different biological systems which can not be obtained by traditional Euclidean-based model. In the first example, we show that hyperbolic representation of odors can better bridge the odor stimuli and the odor response than raw data or the Euclidean representation (Fig.13). There are two future directions: first, this preliminary result needs validation from analysis on larger datasets and requires a deeper understanding of the biological mechanisms; second, the odor perception space is also hyperbolic according to Chapter 1, so how to bridge and unify the whole olfactory system–odor stimuli, odor receptor and odor perception–is an

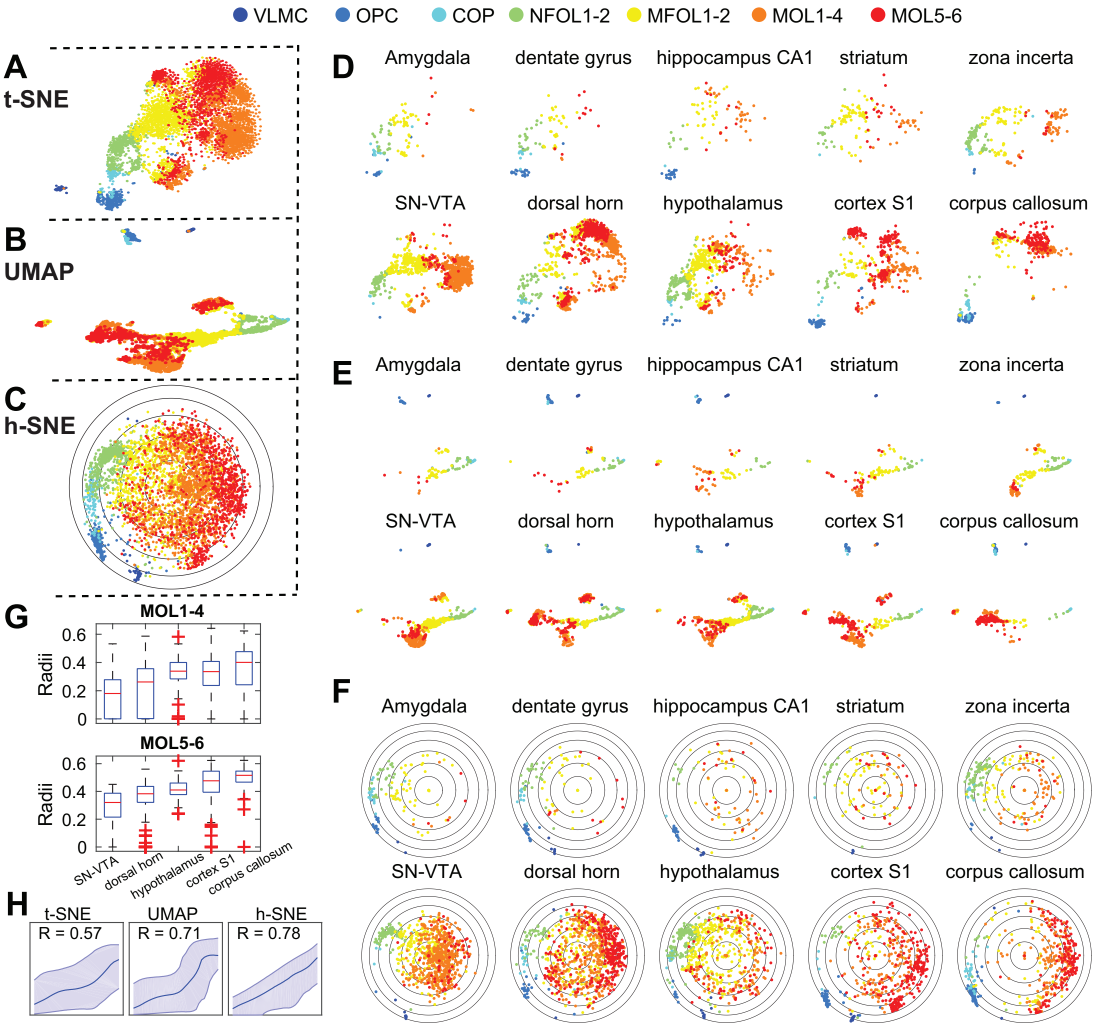

Figure 16. 2D mapping of oligodendrocytes in mouse central nervous system in Marques dataset. The cells are classified the same way as in Marques et al[82], we combine the 13 cells types into seven clusters, the labels are shown at the top of the figure. (A-C) t-SNE, UMAP and h-SNE mapping of oligodendrocytes. (D-F) t-SNE, UMAP and h-SNE mapping of oligodendrocytes in 10 different brain regions. (G) Box plots of radial coordinate distributions of MOL1-4 and MOL5-6 cells in five brain regions. (H)Pairwise distances plots of t-SNE, UMAP and h-SNE embeddings for Marques data, the Pearson correlation coefficients between embedding distances and data distances was shown at the top of the panel.

interesting and important question. In the second example, we use h-SNE to embed the brain cells from nine mouse brain regions, finding that structurally similar cell types display similar spatial localizations regardless of regions (Fig.14). These results provide a novel way to characterize cell types – according to the structures but not according to anatomical regions; of course the validation and explanation of these discoveries still require further works. In the third example, we show that the same mature cell types from different brain regions are quantitatively distinguished by their radii in the hyperbolic disk from h-SNE, which shows an organized region-dependent gene expression patterns of the cells in differentiation. This result seems to contradict with the second example because the same cell types differ in a systematic way in different regions. But from another perspective, the contradiction might be a complement to help us better understand the behaviors of different cell types in dynamic biological systems. These three examples show that hyperbolic representation of biological data may provide new perspectives of data that cannot be revealed by traditional embedding methods, and further study of the biological mechanisms behind them might lead to new discoveries.

As one of the three basic types of geometry, hyperbolic geometry best approximates the complex world because of its equivalence to hierarchical structure. Hopefully many of the other biological systems can also find better accommodations in hyperbolic space, and more biology would be revealed.

Chapter 3, in full, is currently being prepared for submission for publication of the material. Yuansheng Zhou, Tatyana O Sharpee. The dissertation author was the primary investigator and author of this material.

Bibliography

[1] Heidi EM Dobson, Erica M Danielson, and Isaac D Van Wesep. Pollen odor chemicals as modulators of bumble bee foraging on rosa rugosa thunb.(rosaceae). *Plant Species Biology*, 14(2):153–166, 1999.

[2] Anina C Knauer and Florian P Schiestl. Bees use honest floral signals as indicators of reward when visiting flowers. *Ecology letters*, 18(2):135–143, 2015.

[3] Reza Azanchi, Karla R Kaun, and Ulrike Heberlein. Competing dopamine neurons drive oviposition choice for ethanol in drosophila. *Proceedings of the National Academy of Sciences*, 110(52):21153–21158, 2013.

[4] Horace Basil Barlow. Why have multiple cortical areas? *Vision research*, 26(1):81–90, 1986.

[5] Brian Johnson and Ben Shneiderman. Tree-maps: A space-filling approach to the visualization of hierarchical information structures. *Readings in Information Visualization: Using Vision to Think*, pages 152–159, 1999.

[6] Mark A Ragan. Trees and networks before and after darwin. *Biology direct*, 4(1):1–38, 2009.

[7] Dmitri Krioukov, Fragkiskos Papadopoulos, Maksim Kitsak, Amin Vahdat, and Marián Boguná. Hyperbolic geometry of complex networks. *Physical Review E*, 82(3):036106, 2010.

[8] James W Cannon, William J Floyd, Richard Kenyon, Walter R Parry, et al. Hyperbolic geometry. *Flavors of geometry*, 31:59–115, 1997.

[9] Alexei Koulakov, Brian E Kolterman, Armen Enikolopov, and Dmitry Rinberg. In search of the structure of human olfactory space. *Frontiers in systems neuroscience*, 5:65, 2011.

[10] Ernest C Crocker and LF Henderson. *Analysis and classification of odors: an effort to develop a workable method*. 1927.

[11] Chad Giusti, Eva Pastalkova, Carina Curto, and Vladimir Itskov. Clique topology reveals intrinsic geometric structure in neural correlations. *Proceedings of the National Academy of Sciences*, 112(44):13455–13460, 2015.

[12] Michael L Schwieterman, Thomas A Colquhoun, Elizabeth A Jaworski, Linda M Bartoshuk, Jessica L Gilbert, Denise M Tieman, Asli Z Odabasi, Howard R Moskowitz, Kevin M Folta, Harry J Klee, et al. Strawberry flavor: diverse chemical compositions, a seasonal influence, and effects on sensory perception. *PloS one*, 9(2):e88446, 2014.

[13] Denise Tieman, Peter Bliss, Lauren M McIntyre, Adilia Blandon-Ubeda, Dawn Bies, Asli Z Odabasi, Gustavo R Rodríguez, Esther van der Knaap, Mark G Taylor, Charles Goulet, et al. The chemical interactions underlying tomato flavor preferences. *Current Biology*, 22(11):1035–1039, 2012.

[14] Jessica L Gilbert, Matthew J Guthart, Salvador A Gezan, Melissa Pisaroglo de Carvalho, Michael L Schwieterman, Thomas A Colquhoun, Linda M Bartoshuk, Charles A Sims, David G Clark, and James W Olmstead. Identifying breeding priorities for blueberry flavor using biochemical, sensory, and genotype by environment analyses. *PLoS One*, 10(9):e0138494, 2015.

[15] Frank Röck, Karl-Peter Hadeler, Hans-Georg Rammensee, and Peter Overath. Quantitative analysis of mouse urine volatiles: in search of mhc-dependent differences. *PLoS One*, 2(5):e429, 2007.

[16] Ingwer Borg and Patrick JF Groenen. *Modern multidimensional scaling: Theory and applications*. Springer Science & Business Media, 2005.

[17] Rehan M Khan, Chung-Hay Luk, Adeen Flinker, Amit Aggarwal, Hadas Lapid, Rafi Haddad, and Noam Sobel. Predicting odor pleasantness from odorant structure: pleasantness as a reflection of the physical world. *Journal of Neuroscience*, 27(37):10015–10023, 2007.

[18] Manuel Zarzo. Underlying dimensions in the descriptive space of perfumery odors: Part ii. *Food Quality and Preference*, 43:79–87, 2015.

[19] Kobi Snitz, Adi Yablonka, Tali Weiss, Idan Frumin, Rehan M Khan, and Noam Sobel. Predicting odor perceptual similarity from odor structure. *PLoS Comput Biol*, 9(9):e1003184, 2013.

[20] Andrew Dravnieks et al. *Atlas of odor character profiles*. 1985.

[21] Marián Boguná, Fragkiskos Papadopoulos, and Dmitri Krioukov. Sustaining the internet with hyperbolic mapping. *Nature communications*, 1(1):1–8, 2010.

[22] John A Berkowitz and Tatyana O Sharpee. Decoding neural responses with minimal

information loss. *BioRxiv*, page 273854, 2018.

[23] Tarow Indow. *The global structure of visual space*, volume 1. World Scientific, 2004.

[24] Rudolf Karl Luneburg. Mathematical analysis of binocular vision. 1947.

[25] Walter Blumenfeld. The relationship between the optical and haptic construction of space. *Acta Psychologica*, 2:125–174, 1937.

[26] Dmitri Tymoczko. *A geometry of music: Harmony and counterpoint in the extended common practice*. Oxford University Press, 2010.

[27] Kira Belkin, Robyn Martin, Sarah E Kemp, and Avery N Gilbert. Auditory pitch as a perceptual analogue to odor quality. *Psychological Science*, 8(4):340–342, 1997.

[28] Ophelia Deroy, Anne-Sylvie Crisinel, and Charles Spence. Crossmodal correspondences between odors and contingent features: odors, musical notes, and geometrical shapes. *Psychonomic bulletin & review*, 20(5):878–896, 2013.

[29] Avery N Gilbert, Robyn Martin, and Sarah E Kemp. Cross-modal correspondence between vision and olfaction: The color of smells. *The American journal of psychology*, pages 335–351, 1996.

[30] Markus Meister. On the dimensionality of odor space. *Elife*, 4:e07865, 2015.

[31] Caroline Bushdid, Marcelo O Magnasco, Leslie B Vosshall, and Andreas Keller. Humans can discriminate more than 1 trillion olfactory stimuli. *Science*, 343(6177):1370–1372, 2014.

[32] Yilun Zhang and Tatyana O Sharpee. A robust feedforward model of the olfactory system. *PLoS computational biology*, 12(4):e1004850, 2016.

[33] David Zwicker, Arvind Murugan, and Michael P Brenner. Receptor arrays optimized for natural odor statistics. *Proceedings of the National Academy of Sciences*, 113(20):5570–5575, 2016.

[34] Ashok Litwin-Kumar, Kameron Decker Harris, Richard Axel, Haim Sompolinsky, and LF Abbott. Optimal degrees of synaptic connectivity. *Neuron*, 93(5):1153–1164, 2017.

[35] Bin Zhang and Steve Horvath. A general framework for weighted gene co-expression network analysis. *Statistical applications in genetics and molecular biology*, 4(1), 2005.

[36] Wellcome Trust Case Control Consortium et al. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, 447(7145):661, 2007.

[37] Teri A Manolio, Francis S Collins, Nancy J Cox, David B Goldstein, Lucia A Hindorff, David J Hunter, Mark I McCarthy, Erin M Ramos, Lon R Cardon, Aravinda Chakravarti, et al. Finding the missing heritability of complex diseases. *Nature*, 461(7265):747–753, 2009.

[38] Jian Yang, Peter M Visscher, and Naomi R Wray. Sporadic cases are the norm for complex disease. *European Journal of Human Genetics*, 18(9):1039–1043, 2010.

[39] John Novembre, Toby Johnson, Katarzyna Bryc, Zoltán Kutalik, Adam R Boyko, Adam Auton, Amit Indap, Karen S King, Sven Bergmann, Matthew R Nelson, et al. Genes mirror geography within europe. *Nature*, 456(7218):98–101, 2008.

[40] Yuansheng Zhou, Brian H Smith, and Tatyana O Sharpee. Hyperbolic geometry of the olfactory space. *Science advances*, 4(8):eaaq1458, 2018.

[41] A Klimovskaia, D Lopez-Paz, L Bottou, and M Nickel. Poincaré maps for analyzing complex hierarchies in single-cell data. *Nature Communications*, 11(1):2966–2966, 2020.

[42] Jiarui Ding and Aviv Regev. Deep generative model embedding of single-cell rna-seq profiles on hyperspheres and hyperbolic spaces. *BioRxiv*, page 853457, 2019.

[43] Richard C Wilson, Edwin R Hancock, Elżbieta Pekalska, and Robert PW Duin. Spherical and hyperbolic embeddings of data. *IEEE transactions on pattern analysis and machine intelligence*, 36(11):2255–2269, 2014.

[44] Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. In *Advances in neural information processing systems*, pages 6338–6347, 2017.

[45] Jörg A Walter and Helge Ritter. On interactive visualization of high-dimensional data using the hyperbolic plane. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 123–132. ACM, 2002.

[46] Yuval Shavitt and Tomer Tankel. Hyperbolic embedding of internet graph for distance estimation and overlay construction. *IEEE/ACM Transactions on Networking (TON)*, 16(1):25–36, 2008.

[47] Andrej Cvetkovski and Mark Crovella. Low-stress data embedding in the hyperbolic plane using multidimensional scaling. *Appl. Math*, 11(1):5–12, 2017.

[48] Ivan Ovinnikov. Poincaré wasserstein autoencoder. *arXiv preprint arXiv:1901.01427*, 2019.

[49] Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. In

Advances in neural information processing systems, pages 5345–5355, 2018.

[50] Michael B. Eisen, Paul T. Spellman, Patrick O. Brown, and David Botstein. Cluster analysis and display of genome-wide expression patterns. *PNAS; Proceedings of the National Academy of Sciences*, 95(25):14863–14868, 1998.

[51] G. Manning. The protein kinase complement of the human genome. *Science*, 298(5600):1912–1934, Dec 2002.

[52] Erzsébet Ravasz, Anna Lisa Somera, Dale A Mongru, Zoltán N Oltvai, and A-L Barabási. Hierarchical organization of modularity in metabolic networks. *science*, 297(5586):1551–1555, 2002.

[53] Andreas Dunkel, Martin Steinhaus, Matthias Kotthoff, Bettina Nowak, Dietmar Krautwurst, Peter Schieberle, and Thomas Hofmann. Nature's chemical signatures in human olfaction: a foodborne perspective for future biotechnology. *Angewandte Chemie International Edition*, 53(28):7124–7143, 2014.

[54] David Meunier, Renaud Lambiotte, Alex Fornito, Karen Ersche, and Edward T Bullmore. Hierarchical modularity in human brain functional networks. *Frontiers in neuroinformatics*, 3:37, 2009.

[55] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA, 1967.

[56] Sam T Roweis and Lawrence K Saul. Nonlinear dimensionality reduction by locally linear embedding. *science*, 290(5500):2323–2326, 2000.

[57] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(Nov):2579–2605, 2008.

[58] Yuansheng Zhou and Tatyana Sharpee. Using global t-sne to preserve inter-cluster data structure. *bioRxiv*, page 331611, 2018.

[59] Etienne Becht, Leland McInnes, John Healy, Charles-Antoine Dutertre, Immanuel WH Kwok, Lai Guan Ng, Florent Ginhoux, and Evan W Newell. Dimensionality reduction for visualizing single-cell data using umap. *Nature biotechnology*, 37(1):38, 2019.

[60] Christopher De Sa, Albert Gu, Christopher Ré, and Frederic Sala. Representation tradeoffs for hyperbolic embeddings. *Proceedings of machine learning research*, 80:4460, 2018.

[61] Roger N Shepard. Multidimensional scaling, tree-fitting, and clustering. *Science*, 210(4468):390–398, 1980.

[62] Margus Lukk, Misha Kapushesky, Janne Nikkilä, Helen Parkinson, Angela Goncalves, Wolfgang Huber, Esko Ukkonen, and Alvis Brazma. A global map of human gene expression. *Nature biotechnology*, 28(4):322, 2010.

[63] Michael B Elowitz, Arnold J Levine, Eric D Siggia, and Peter S Swain. Stochastic gene expression in a single cell. *Science*, 297(5584):1183–1186, 2002.

[64] Marjorie F Oleksiak, Gary A Churchill, and Douglas L Crawford. Variation in gene expression within and among natural populations. *Nature genetics*, 32(2):261, 2002.

[65] Arjun Raj and Alexander van Oudenaarden. Nature, nurture, or chance: stochastic gene expression and its consequences. *Cell*, 135(2):216–226, 2008.

[66] GTEx Consortium et al. Genetic effects on gene expression across human tissues. *Nature*, 550(7675):204, 2017.

[67] Xiaoping Han, Renying Wang, Yincong Zhou, Lijiang Fei, Huiyu Sun, Shujing Lai, Assieh Saadatpour, Ziming Zhou, Haide Chen, Fang Ye, et al. Mapping the mouse cell atlas by microwell-seq. *Cell*, 172(5):1091–1107, 2018.

[68] Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161(5):1202–1214, 2015.

[69] Andrew Butler, Paul Hoffman, Peter Smibert, Efthymia Papalexi, and Rahul Satija. Integrating single-cell transcriptomic data across different conditions, technologies, and species. *Nature biotechnology*, 36(5):411–420, 2018.

[70] Peter J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53 – 65, 1987.

[71] Sylvie Deborde, Tatiana Omelchenko, Anna Lyubchik, Yi Zhou, Shizhi He, William F McNamara, Natalya Chernichenko, Sei-Young Lee, Fernando Barajas, Chun-Hao Chen, et al. Schwann cells induce cancer cell dispersion and invasion. *The Journal of clinical investigation*, 126(4):1538–1554, 2016.

[72] Bàrbara Castellana, Daniel Escuin, Gloria Peiró, Bárbara Garcia-Valdecasas, Tania Vázquez, Cristina Pons, Maitane Pérez-Olabarria, Agustí Barnadas, and Enrique Lerma. Aspn and gjb2 are implicated in the mechanisms of invasion of ductal breast carcinomas. *Journal of Cancer*, 3:175, 2012.

[73] Ye Song, Shihao Zheng, Jizhou Wang, Hao Long, Luxiong Fang, Gang Wang, Zhiyong Li, Tianshi Que, Yi Liu, Yilei Li, et al. Hypoxia-induced plod2 promotes proliferation,

migration and invasion via pi3k/akt signaling in glioma. *Oncotarget*, 8(26):41947, 2017.

[74] Victoria Moignard, Steven Woodhouse, Laleh Haghverdi, Andrew J Lilly, Yosuke Tanaka, Adam C Wilkinson, Florian Buettner, Iain C Macaulay, Wajid Jawaid, Evangelia Diamanti, Shin-Ichi Nishikawa, Nir Piterman, Valerie Kouskoff, Fabian J Theis, Jasmin Fisher, and Berthold Göttgens. Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nature Biotechnology*, 33(3):269–276, feb 2015.

[75] Yan Wu, Pablo Tamayo, and Kun Zhang. Visualizing and interpreting single-cell gene expression datasets with similarity weighted nonnegative embedding. *Cell systems*, 7(6):656–666, 2018.

[76] Jiarui Ding, Anne Condon, and Sohrab P Shah. Interpretable dimensionality reduction of single cell transcriptome data with deep generative models. *Nature communications*, 9(1):2002, 2018.

[77] Dmitry Kobak and Philipp Berens. The art of using t-sne for single-cell transcriptomics. *bioRxiv*, page 453449, 2018.

[78] Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 1(10):e2, 2016.

[79] Joseph B Kruskal. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1–27, 1964.

[80] Elissa A Hallem and John R Carlson. Coding of odors by a receptor repertoire. *Cell*, 125(1):143–160, 2006.

[81] Arpiar Saunders, Evan Z Macosko, Alec Wysoker, Melissa Goldman, Fenna M Krienen, Heather de Rivera, Elizabeth Bien, Matthew Baum, Laura Bortolin, Shuyu Wang, et al. Molecular diversity and specializations among the cells of the adult mouse brain. *Cell*, 174(4):1015–1030, 2018.

[82] Sueli Marques, Amit Zeisel, Simone Codeluppi, David van Bruggen, Ana Mendanha Falcão, Lin Xiao, Huiliang Li, Martin Häring, Hannah Hochgerner, Roman A Romanov, et al. Oligodendrocyte heterogeneity in the mouse juvenile and adult central nervous system. *Science*, 352(6291):1326–1329, 2016.