Uncovering Developmental Lineages from Single-cell Data with Contrastive Poincaré Maps

Nithya Bhasker[1] nithya.bhasker@nct-dresden.de
Hattie Chung[2] hattie.chung@yale.edu
Louis Boucherie[3] louibo@dtu.dk
Vladislav Kim[4] vladislav.kim@bayer.com
Stefanie Speidel[1] stefanie.speidel@nct-dresden.de
Melanie Weber[5] mweber@seas.harvard.edu

[1] *Department of Translational Surgical Oncology, National Center for Tumor Diseases (NCT/UCC Dresden), Dresden, Germany: German Cancer Research Center (DKFZ), Heidelberg, Germany; Faculty of Medicine and University Hospital Carl Gustav Carus, TUD Dresden University of Technology, Dresden, Germany; Helmholtz-Zentrum Dresden-Rossendorf (HZDR), Dresden, Germany*

[2] *Department of Medicine, Cardiovascular Research Center, Yale School of Medicine*

[3] *DTU Compute, Technical University of Denmark; Center for Social Data Science, University of Copenhagen*

[4] *Machine Learning Research, Bayer AG, Berlin, Germany*

[5] *Harvard University, School of Engineering and Applied Sciences*

Abstract

Embeddings play a central role in single-cell RNA sequencing (scRNA-seq) data analysis by transforming complex gene expression profiles into interpretable, low-dimensional representations. While Euclidean embeddings distort hierarchical relationships in low dimensions, hyperbolic geometry can represent hierarchies accuractely in low dimensions. However, existing hyperbolic methods, such as Poincaré Maps (PM), lose accuracy in deeper hierachies and require extensive feature engineering and memory. We present Contrastive Poincaré Maps (CPM), a scalable approach that reliably preserves inherent hierarchical structures. On synthetic trees with up to five generations and 34,000 individuals, CPM reduces distortion by 99% (1.9 vs. 126.3) and requires 13-fold less memory than PM. We demonstrate CPM's utility across three case studies: scalable analysis of 116,312 mouse gastrulation cells, accurate reconstruction of hierarchical structure in mouse hematopoiesis, and faithful representation of multi-lineage hierarchies in chicken cardiogenesis. By integrating hyperbolic geometry with contrastive learning, CPM enables scalable, structure-preserving embeddings for developmental scRNA-seq data. Code: https://github.com/NithyaBhasker/ContrastivePoincareMaps

1 Introduction

Embedding methods play a central role in single-cell RNA sequencing (scRNA-seq) data analysis by transforming complex gene expression profiles into interpretable, low-dimensional representations. Advances in high-throughput scRNA-seq have transformed our ability to study dynamic

A preliminary version of a part of this work was presented at the ICLR Workshop on Machine Learning for Genomics Explorations Bhasker et al. (2024).

biological processes by profiling gene expression at single-cell resolution Macosko et al. (2015); Tang et al. (2009). It has been particularly powerful for dissecting lineage-specific information and differentiation processes during development Treutlein et al. (2014); Trapnell et al. (2014); Paul et al. (2015); Nestorowa et al. (2016); La Manno et al. (2018a); Pijuan-Sala et al. (2019). The resulting datasets capture thousands of genes across tens to hundreds of thousands of cells, making them inherently complex and high-dimensional. Computational tools extract biologically meaningful patterns from this complexity, such as the branching of cells into distinct lineages or transitions between cell states.

Dimensionality reduction methods like UMAP McInnes et al. (2018), t-SNE Van der Maaten and Hinton (2008), and MDS Kruskal and Wish (1978) are widely used to visualize cellular relationships in low-dimensional spaces. Additionally, specialized algorithms for single-cell data, including PHATE Moon et al. (2019), diffusion maps Nadler et al. (2005), Monocle Trapnell et al. (2014), and PAGA Wolf et al. (2019), enable reconstructing developmental trajectories, estimating pseudotime, and inferring gene expression dynamics, critical for insights into cell fate decisions and uncovering the regulatory programs that drive tissue development and cellular diversity. A common feature of both general-purpose and single-cell-specific computational tools is their reliance on Euclidean embeddings to represent high-dimensional gene expression data in two or three dimensions for visualization and analysis. However, Euclidean space has been found to be inherently limited in capturing hierarchical or tree-like relationships, such as those underlying cell differentiation and lineage branching during development, leading to distortions in representing developmental trajectories, particularly when complex bifurcations or nested lineage hierarchies are present Nickel and Kiela (2017).

To overcome this limitation, hyperbolic spaces, which are negatively curved geometric manifolds, have been proposed as a powerful alternative, as they are mathematically well-suited for embedding hierarchical structures with high fidelity even in low dimensions Nickel and Kiela (2017); Sala et al. (2018).Building on this insight, Poincaré Maps (PM) Klimovskaia et al. (2020) was introduced to embed single-cell data into hyperbolic space. This approach extends a broader line of research on hyperbolic graph embeddings Nickel and Kiela (2017); Chamberlain et al. (2017); Chami et al. (2019) and their advantages in downstream machine learning tasks Ganea et al. (2018); Weber et al. (2020). Poincaré Maps has demonstrated the ability to learn biologically meaningful representations across several single-cell datasets Klimovskaia et al. (2020). However, several practical limitations have been identified. Although PM can effectively represent data structured as shallow hierarchies, its performance degrades with increasing tree depth, making it less suitable for modeling complex developmental trajectories Bhasker et al. (2024). In addition, the method depends on manual feature engineering to achieve good performance, entails a memory-intensive training process, and lacks support for efficient updates as new data become available Klimovskaia et al. (2020); Bhasker et al. (2024). These factors limit its scalability and usability in large-scale single-cell studies.

Euclidean embedding methods introduce distortions in the representation of developmental trajectories, whereas the hyperbolic Poincaré Maps approach faces computational and scalability constraints. To address these limitations, we present Contrastive Poincaré maps (CPM), a scalable contrastive learning-based approach for efficiently representing high-dimensional single-cell data in a low-dimensional hyperbolic space while preserving the inherent hierarchical structures in the data. Our method builds on PM Klimovskaia et al. (2020), as well as contrastive approaches for representing tabular data Bahri et al. (2022). We validate our method using syn-

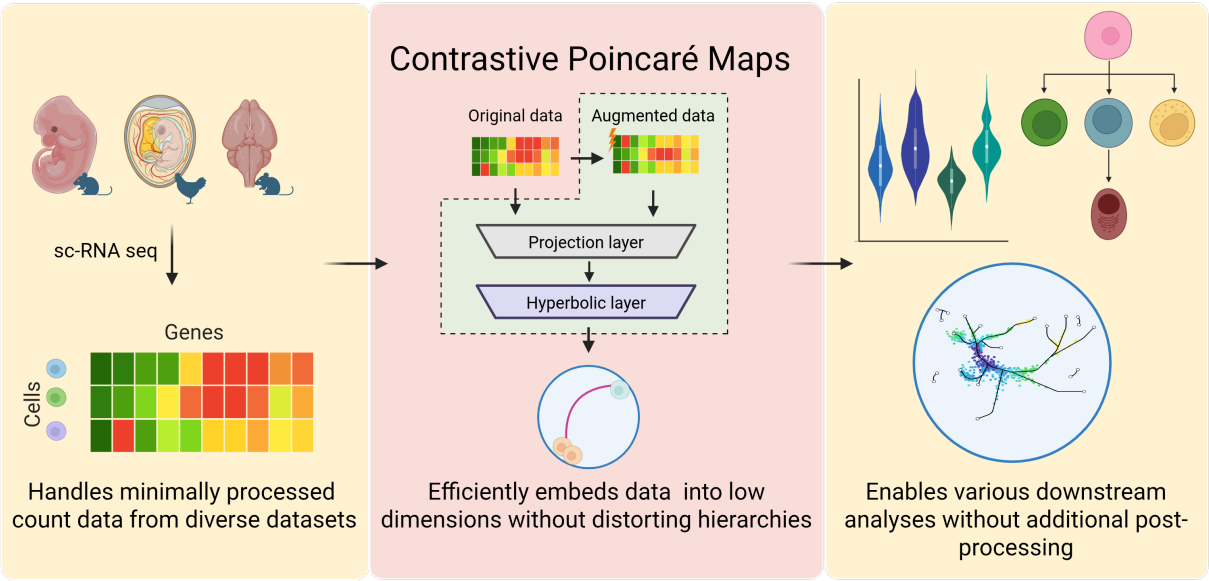

Figure 1: **Contrastive Poincaré Maps** is an efficient, low-distortion embedding method that handles minimally processed transcriptomic data from a wide range of specimens and supports multiple downstream analyses without further post-processing, delivering especially strong performance in lineage detection applications. Figure was created in BioRender. Bhasker, N. (2025) https://BioRender.com/t83d482

thetic data and three published scRNA-seq datasets. On synthetic trees with up to 5 generations and 34,000 individuals, CPM cuts distortion by 99% (1.9 vs. 126.3) and requires 13-fold less memory relative to PM. CPM uncovers accurate hierarchies across 9 developmental stages in the mouse gastrulation dataset comprising 116,312 cells, disentangles global multi-lineage hierarchies in the chicken cardiogenesis dataset while preserving intra-lineage developmental trends, and enables sampling-density-invariant hierarchical analysis in the mouse hematopoiesis dataset. By integrating hyperbolic geometry with contrastive learning, CPM delivers a scalable embedding framework that preserves hierarchical dependencies in developmental lineages, accelerates exploratory data analysis, and opens new avenues for biological discovery in developmental processes using scRNA-seq data.

2 Results

We present Contrastive Poincaré maps (CPM), a contrastive learning-based approach to represent high-dimensional single cell data in a low-dimensional hyperbolic space (see Fig. 1). Our method builds on prior work on shallow hyperbolic embedding methods for single cell data analysis Klimovskaia et al. (2020), as well as contrastive approaches for representing tabular data Bahri et al. (2022). Contrastive learning facilitates improved data representation by effectively managing complexity in higher dimensions Chen et al. (2020), demonstrating proven feature representation capabilities Grill et al. (2020), robustness to noise Ghosh and Lan (2021), and generalization across various downstream tasks Oord et al. (2018); Radford et al. (2021).

2.1 Overview of the Framework and Experimental setup

The input to the framework is a count matrix in the form of cells $\times$ genes. CPM processes the gene expression matrix in batches, applying an augmentation strategy that randomly replaces gene expression values with those from random cells in the dataset. This kind of augmentation strategy contributes to the learning of robust representations across diverse cells. We subsequently project both views of the data into a latent space. The function of the projection block is twofold, acting as both a dimensionality reduction and noise reduction method, which can be adapted and trained end-to-end, unlike conventional methods such as Principal Component Analysis (PCA). The latent representations of the data are then embedded into hyperbolic space using hyperbolic graph neural networks Chami et al. (2019). We choose a two-dimensional representation space, the Poincaré disk, with curvature $\kappa = -1$.

The choice of a two-dimensional representation space is motivated by ease of visualization and interpretability of the resulting embeddings; varying or learning the curvature as opposed to using a fixed value has been shown to have little effect (see, e.g., Chami et al. (2019)), motivating our choice of $\kappa = -1$. We use the classical InfoNCE contrastive loss for training our model. The distance between both views of the data is calculated in hyperbolic space. The embeddings are learned by reducing the distance between similar cells and increasing the distance between dissimilar cells. Much like any deep learning framework, CPM operates with a small set of tunable hyperparameters. The number of projection and hyperbolic layers, rate of corruption for the augmentation strategy, batch size and latent space dimension are some of the key hyperparameters of our model. The model maintains stable performance across a broad range of these hyperparameter settings, demonstrating robustness to hyperparameter variation (see Supplementary Fig. 11). The number of projection layers and the corruption rate require a moderate balance, to achieve embeddings with minimum distortion (see Fig. 11).

To evaluate CPM, we conducted experiments with multiple datasets: deep synthetic binary trees Moon et al. (2019), early organogenesis in mouse gastrulation Pijuan-Sala et al. (2019), developmental trajectories in mouse hematopoiesis Paul et al. (2015), and multi-lineage developmental trajectories in chicken cardiogenesis Mantri et al. (2021).We benchmarked CPM against widely used dimensionality reduction, visualization, and lineage detection methods, including t-SNE Maaten and Hinton (2008), UMAP McInnes et al. (2018), and PHATE Moon et al. (2019). We also compared our model against Poincaré Maps (PM), which has outperformed the aforementioned widely used approaches on both synthetic and real world tree-like data.

Evaluation was performed using two standard metrics from manifold embedding literature: Q-scores Lee and Verleysen (2010) and worst-case distortion (see Sec. 4.7). Q-scores offer an unsupervised, scale-free, rank-invariant criterion for assessing embeddings. By operating on neighbor ranks rather than raw pairwise distances, they remain unaffected by global scaling, making them a robust baseline metric for evaluating the fidelity of any dimensionality-reduction method. Worst-case distortion provides a provable upper bound on the weakest links in the embedding, making downstream tasks like lineage detection more trustworthy.

2.2 Enhanced global accuracy and efficiency of CPM over Poincaré Maps

Quantitative evaluation and benchmarking on synthetic data are essential for validating embedding methods because they enable objective, reproducible assessment of how well the embeddings recover known ground-truth structures, providing a controlled basis for comparing

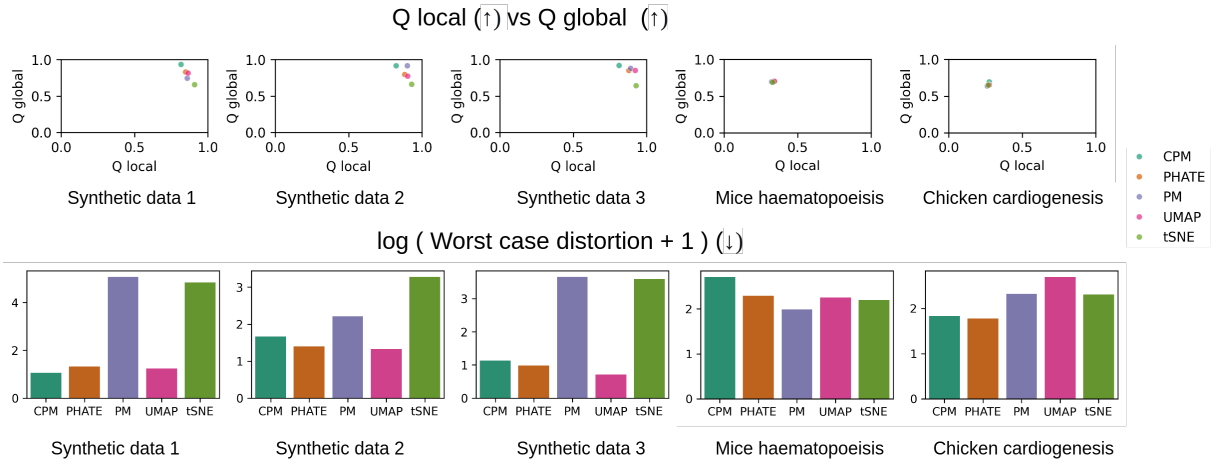

Figure 2: **Quantitative evaluation**. Contrastive Poincaré Maps achieves competitive representation accuracy across diverse real and synthetic datasets, requiring minimal hyperparameter tuning and no additional PCA-based dimensionality reduction. Synthetic data 1 and Synthetic data 2 represent deep (15 and 10 hierarchical levels, respectively), heterogeneous, balanced binary trees, while Synthetic data 3 is a deep, heterogeneous, imbalanced binary tree. Real datasets encompass shallow hierarchies (mice haematopoiesis and mouse brain) and a shallow, multi-lineage topology (chicken cardiogenesis).

performance, and scalability across methods. To assess and compare the representation accuracy and computational efficiency of our model, we generated a diverse collection of synthetic datasets varying in size and structural complexity Moon et al. (2019). For illustrative purposes, we selected three representative examples: (i) **Synthetic data 1** – a deep binary tree with 15 hierarchical levels, where each node branches into two children; (ii) **Synthetic data 2** – an imbalanced binary tree with 10 levels, in which each node also has two children, but branches are sampled randomly to introduce structural irregularity; and (iii) **Synthetic data 3** – a balanced binary tree with 10 levels, each node branching into two children, with no randomness in structure. To enhance heterogeneity within the trees, one branch at each depth level was constrained to contain one third of the total samples for that level.

CPM achieves high global accuracy across all datasets with low distortion (Fig. 2). Compared to Euclidean state-of-the-art methods, CPM attains comparable or superior performance with minimal hyperparameter tuning (Supplementary Fig. 11), despite not relying on additional preprocessing steps such as Principal Component Analysis (PCA) or batch effect correction with Scanorama Hie et al. (2024). For a fair comparison, Supplementary Fig. 10 shows the performance of all methods without any additional preprocessing. On closer inspection, visualizations of the embeddings produced by PHATE, t-SNE, and UMAP display unnatural intersections between branches, distorting the expected developmental trajectories (see Supplementary Figures 12, 13, 14).

When compared to the hyperbolic baseline, Poincaré Maps, CPM substantially improves both representational accuracy and computational efficiency. On synthetic trees with up to five generations and 34,000 individuals, CPM reduces distortion by 99% (1.9 vs. 126.3) and requires 13-fold less memory relative to PM (Fig. 3). Notably, in terms of memory usage, CPM

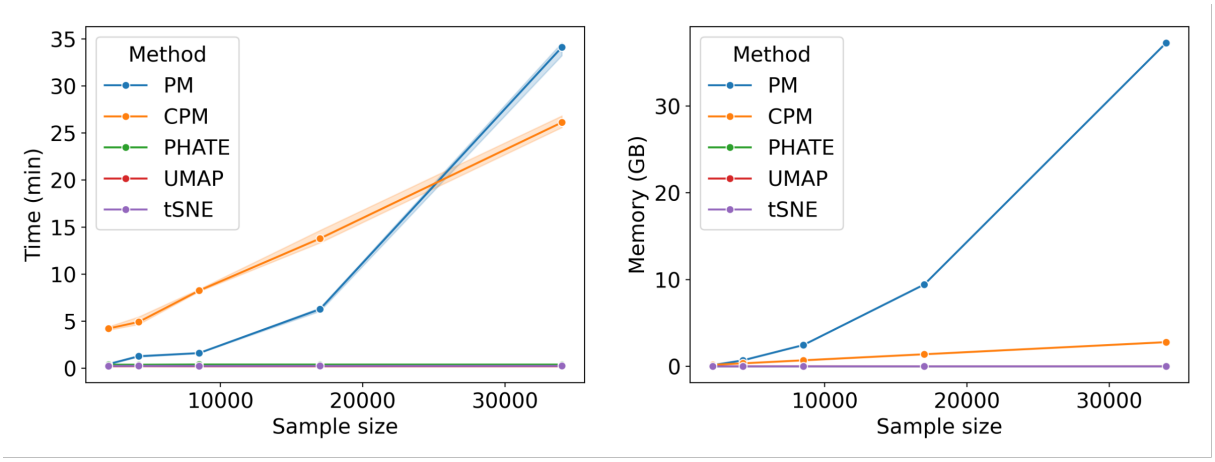

Figure 3: **Space and time complexity analysis**. Time and memory requirements for each embedding approach are assessed across varying sample sizes using synthetic data. Contrastive Poincaré Maps (CPM) demonstrates substantial efficiency improvements in both computational time and memory usage compared to data-driven counterparts requiring explicit training steps.

remains competitive with shallow, untrained methods that have a far smaller parameter budget. Visualizations of embeddings across individual datasets further highlight the differences (Supplementary Figures 12, 13, 14). In contrast to CPM, Poincaré Maps fail to preserve continuous progression between hierarchical levels: all levels beyond the root are compressed toward the horizon, producing unnatural representations that distort the true hierarchical organization (Supplementary Fig. 12).

In summary, CPM not only matches Euclidean state-of-the-art methods in accuracy but also surpasses Poincaré Maps in both representational accuracy and computational efficiency, establishing it as a powerful and scalable framework for single-cell embedding.

2.3 Scalable Analysis of Mouse Gastrulation and Early Organogenesis

The mouse gastrulation and early organogenesis dataset Pijuan-Sala et al. (2019), encompasses 116,312 cells from mouse embryos across nine developmental stages between embryonic days 6.5 and 8.5. This dataset captures transcriptional dynamics during the critical phases of gastrulation and early organ formation. To demonstrate the scalability and ability of CPM to capture these complex developmental hierarchies at large scale, we performed downstream analysis on the embedding of the dataset. The dataset was normalized to 10000 counts per cell, Log1p transformed and filtered to contain 2000 highly variable genes. The first important observation is that state-of-the-art approaches, except CPM, fail to produce an embedding for the complete dataset (containing 100,000 cells), due to their reliance on pairwise distances for the computation of embeddings, which scales quadratically in the number of cells and runs out of memory on a fairly sophisticated infrastructure (cluster equipped with 112 CPUs with 1TB of RAM). Since our method processes data in batches during training, it efficiently generates embeddings within a reasonable time (approx. 2-3 hours). Unfortunately, the metrics that we use in our work also rely on pairwise distances for their computation and fail to produce results for the complete dataset. However, the performance of all the methods on a subsample of the dataset is illustrated

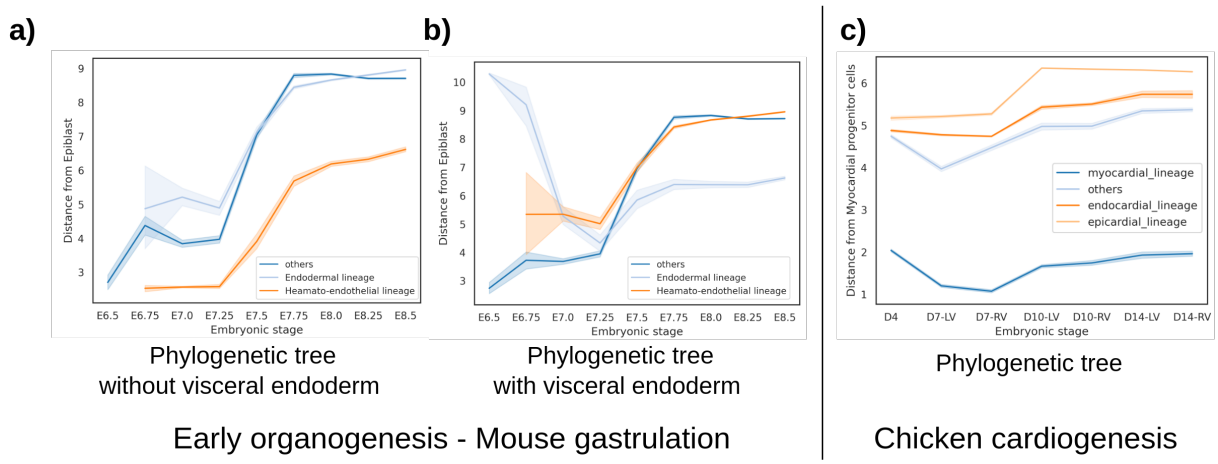

Figure 4: **Application to organogenesis**. Illustration of how distances within Contrastive Poincaré Maps reflect developmental progression across lineages in mouse gastrulation and chicken cardiogenesis datasets. (a, b) Differentiation across embryonic stages in the endodermal and haemato-endothelial lineages from the mouse gastrulation dataset: (a) Endodermal lineage excluding visceral endoderm; (b) Endodermal lineage including visceral endoderm, supported by evidence for the presence of visceral endoderm as early as embryonic stage E6.5. (c) Differentiation across developmental stages in myocardial, epicardial, and endocardial lineages from the chicken cardiogenesis dataset.

in Supplementary Fig 18.

In line with the findings of Pijuan-Sala et al. (2019), we observe (based on Figs. 4, 5, 6): (i) a decline in the frequency of pluripotent epiblast cells over time, (ii) the appearance of mesodermal and definitive endodermal cells as early as E6.75, (iii) the origin of gut endoderm from visceral as well as definitive endoderm cells, (iv) the formation of red blood cells in two consecutive waves — the first at E7.5 and the second at E8.25, and (v) high expression of Kdr gene in the endothelial region (see Supplementary Fig 25).

A simple phylogenetic tree, as seen in Fig. 4, reveals the presence of mature visceral endoderm cells as early as E6.5. This could be explained by the fact that the authors sampled extra-embryonic structures alongside the gastrulating embryo to investigate the convergence of primitive streak-derived definitive endoderm with visceral endoderm-derived cells at the molecular level. The original study Pijuan-Sala et al. (2019) performed analyses such as flow cytometry, in addition to scRNA-seq, to arrive at these findings. In contrast, CPM facilitated the observation of these findings with a single embedding method with minimal hyperparameter tuning.

One of the main conclusions of Pijuan-Sala et al. Pijuan-Sala et al. (2019) was that densely sampled large-scale single-cell profiling has the potential to help advance the understanding of embryonic development. Thanks to its scalability, CPM enables such large-scale analyses, whereas other embedding methods are prohibitively memory-intensive in this regime.

2.4 Accurate Hierarchical Representation in Mice Hematopoiesis

The mice hematopoiesis dataset Paul et al. (2015) comprises of scRNA-seq profiles of approximately 2,730 mouse bone marrow-derived myeloid progenitor cells, including common myeloid progenitors (CMP), granulocyte-monocyte progenitors (GMP), and megakaryocyte–erythroid

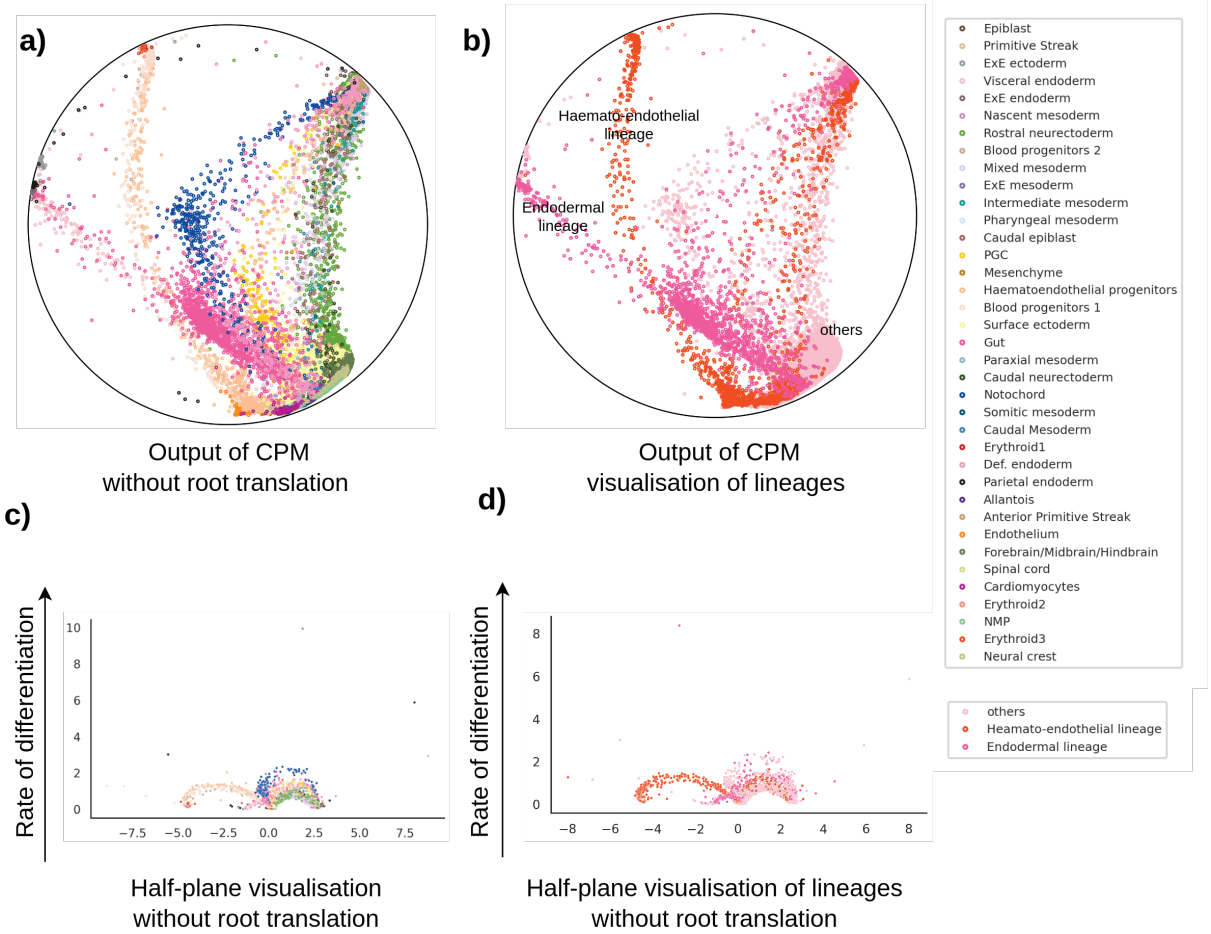

Figure 5: **Analysis of early organogenesis during mouse gastrulation.** (a, b) Visualization of the embedding learned by Contrastive Poincaré Maps for the mouse gastrulation dataset, comprising 116,312 cells across nine developmental stages. Contrastive Poincaré Maps accurately preserves the global hierarchy of lineages (b) as well as the relationships among individual cells (a). (c, d) Visualization of the embedding following transformation into the Poincaré half-plane, enabling inference of differentiation origins and rates.

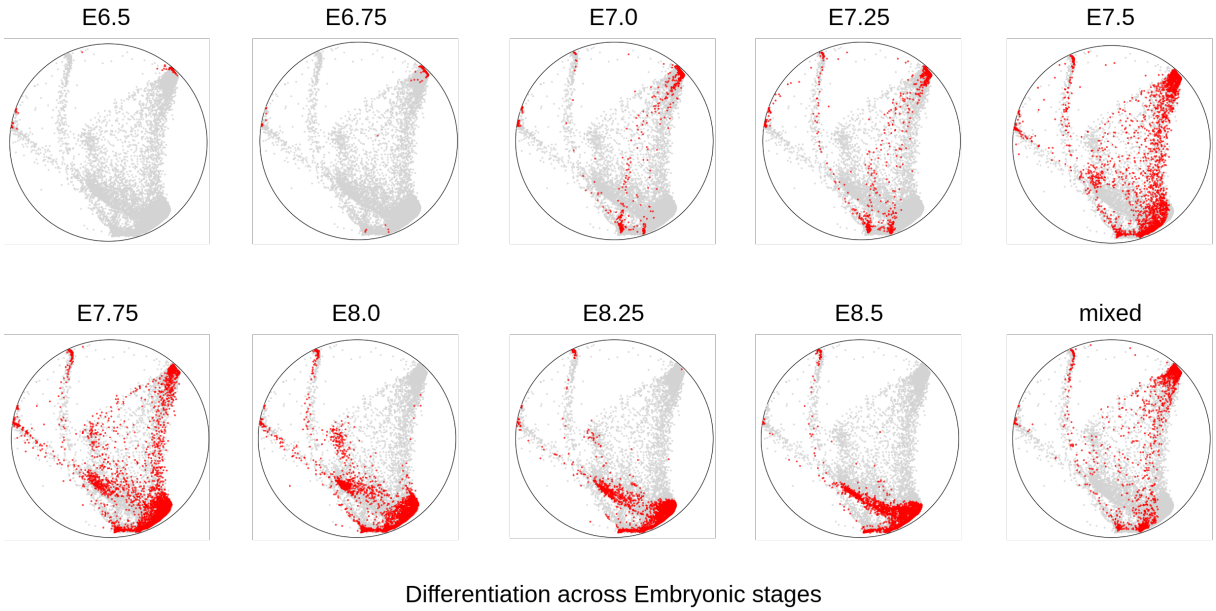

Figure 6: **Early organogenesis during mouse gastrulation** Time-lapse visualization of cell differentiation across nine developmental stages overlaid on the embedding produced by Contrastive Poincaré Maps. The embedding provides evidence for the early emergence of endothelial lineage cells and highlights two distinct waves in the formation of red blood cells.

progenitor (MEP). It captures transcriptional heterogeneity across these cell types, revealing lineage-specific gene expression patterns. This dataset is widely used to study hematopoietic differentiation, transcriptional priming, and to benchmark single-cell analysis methods. The mice hematopoiesis data Paul et al. (2015) revealed that GMPs are transcriptionally heterogeneous and already primed for distinct myeloid fates (neutrophils, monocytes, and basophils), challenging the classical stepwise model of hematopoiesis. The transcription factor Cebpa was identified as a key regulator in this early lineage commitment. Regarding basophil origin, the authors presented strong evidence for a basophil-neutrophil-monocyte progenitor (BNMP) within the GMP pool, supporting a model of gradual, transcriptionally guided differentiation.

However, subsequent analyses using Partition-based Graph Abstraction (PAGA) Wolf et al. (2019) suggested that basophil lineage trajectories are dataset-dependent and may reflect multiple origins. While mice hematopoiesis data Paul et al. (2015) supported the BNMP model, other datasets such as Nestorowa et al. Nestorowa et al. (2016) suggested erythroid-megakaryocyte-basophil progenitor model, and the more densely sampled Dahlin et al. Dahlin et al. (2018) dataset provided evidence for both possibilities. Additionally, it was observed that Poincaré maps Klimovskaia et al. (2020) effectively captured the hierarchical relationships among various myeloid progenitor populations. Notably, Poincaré maps positioned the 16Neu cluster downstream of the 15Mo cluster, suggesting a potential sequential relationship between these cell types. This contrasts with the canonical hematopoietic hierarchy, where neutrophils and monocytes are typically considered to arise independently from a common progenitor. This discrepancy was attributed to a possible mislabeling in the original dataset Klimovskaia et al. (2020).

This indicates that the development process in mice hematopoiesis may not follow a singular, fixed path, but rather involve context-dependent or plastic differentiation routes. Moreover, the sampling density of the datasets gave rise to different conclusions. To enable distortion-free analysis of the developmental lineages, while being invariant to sampling density, we used CPM to analyze the developmental lineages of the mice hematopoiesis dataset. We obtained the processed dataset from Klimovskaia et al. (2020). More details on the dataset processing and the hype parameters used for the CPM model is provided in Section 4.

Fig. 7 illustrates the canonical cell ordering derived from the embeddings generated from CPM, PM and PHATE. For each method, we plot the median distance of cells from the root against their angular separation. Among the methods, only CPM successfully distinguishes MEP from GMP, thereby supporting the BNMP model. This is consistent with both the dataset characteristics and previous findings using PAGA. Furthermore, in contrast to PM and PHATE, CPM positions dendritic cells proximal to eosinophils and downstream of the lymphoid lineage, rather than near erythroid cells. This arrangement aligns with established hierarchies (Fig. 7 a)).

We also visualize the trajectory between the 1Ery and 14Mo cell types by interpolating between between points along these branches. For CPM and PM, interpolation is performed along geodesics (see Appendix A.4.2). In CPM embedding, the trajectory passes through intermediate Ery and Mk cell types, consistent with known biological progression (Fig. 8). In contrast, this pattern is not observed in PM or PHATE embeddings. Visualizations of embeddings and gene expressions for certain genes of interest are provided in Supplementary Figures 16, 26.

These results indicate that CPM not only accurately represents the individual cell types, but also captures the underlying structure of the dataset by preserving meaningful transitional relationships in the embedding space.

2.5 Multi-Lineage Hierarchical Representation in Chicken Cardiogenesis

The chicken cardiogenesis dataset Mantri et al. (2021) includes transcriptomic profiles from 22,315 individual cells collected across four embryonic stages (days 4, 7, 10, and 14). This dataset contains gene expression profiles from cells spanning multiple lineages and provides a high-resolution view of how cellular differentiation and morphogenetic processes are coordinated during heart development, offering key insights into cardiac lineage specification. To demonstrate the ability of CPM to capture global hierarchies across multiple lineages, we analyzed the chicken cardiogenesis dataset. The dataset was processed following the approach described in Mantri et al. (2021), with the exception that CPM operated on the full dataset containing 18,797 genes prior to Scanorama integration and the selection of the top 2000 highly variable genes.

State of the art methods achieve optimal performance primarily when additional preprocessing steps, such as Scanorama integration, were applied to correct for batch effects in the dataset. In contrast, our approach distinctly clusters the myocardial, endocardial, and epicardial lineages without requiring such corrections, while simultaneously preserving the intrinsic global hierarchical structure of the data. Notably, our method also clearly separates hematopoiesis related cell populations, including erythrocytes, macrophages, and dendritic cells (see Fig. 9a).

To further analyze lineage organization using CPM, we trained embeddings for each lineage independently (see Fig.9c). These embeddings reveal well preserved intra-lineage hierarchies and effectively capture the differentiation continuum from immature to mature cell states across developmental stages (see Fig.9d). The gene expression for certain genes of interest and the visu-

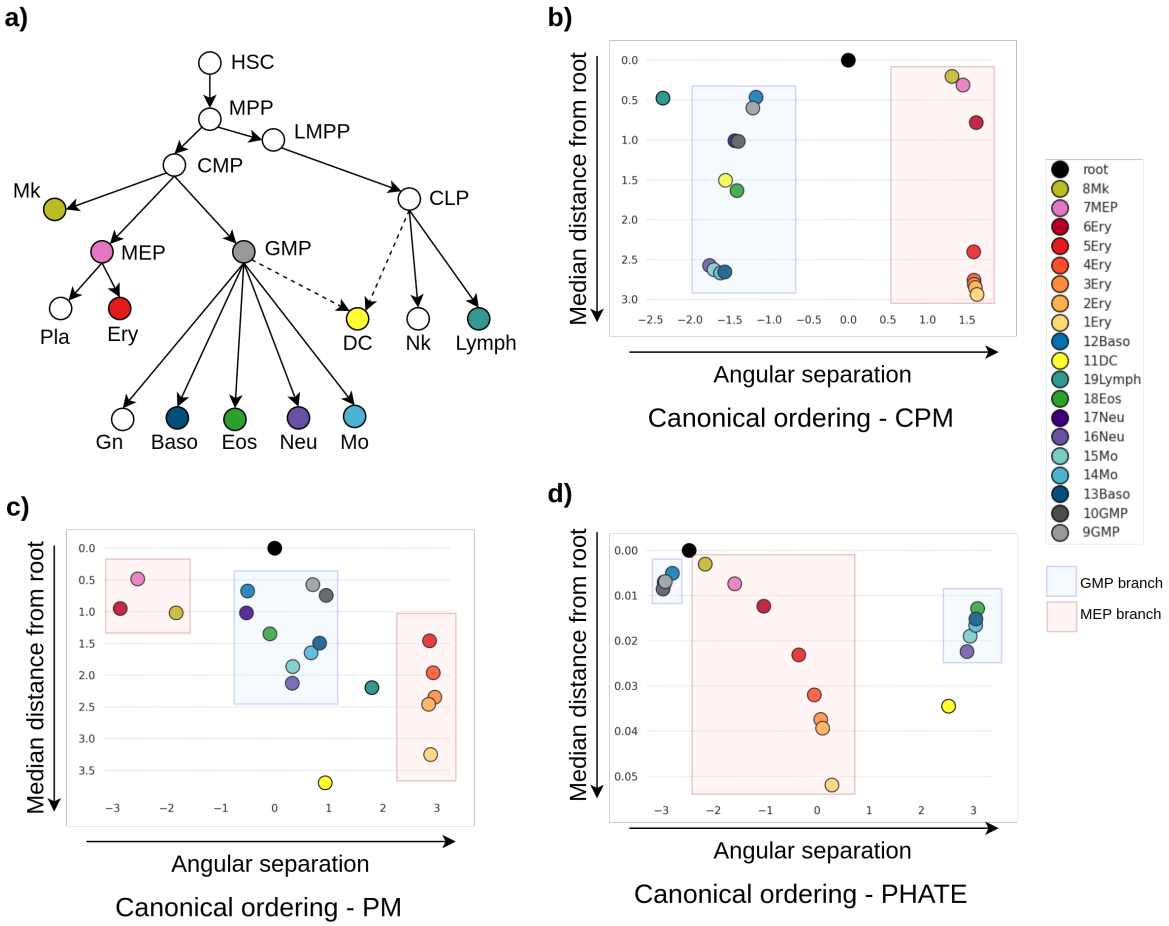

Figure 7: **Canonical ordering of cell types** Distances and angular relationships in the Contrastive Poncaré Maps embedding more faithfully preserve the canonical haematopoietic hierarchy than Poincaré Maps or PHATE, with particularly accurate reconstruction of the myeloid branch comprising basophils, neutrophils, monocytes, and eosinophils. (a) Reference haematopoietic lineage tree. (b–d) Cell type ordering based on pairwise distances and angles in Contrastive Poincaré Maps, Poincaré Maps, and PHATE embeddings, respectively.

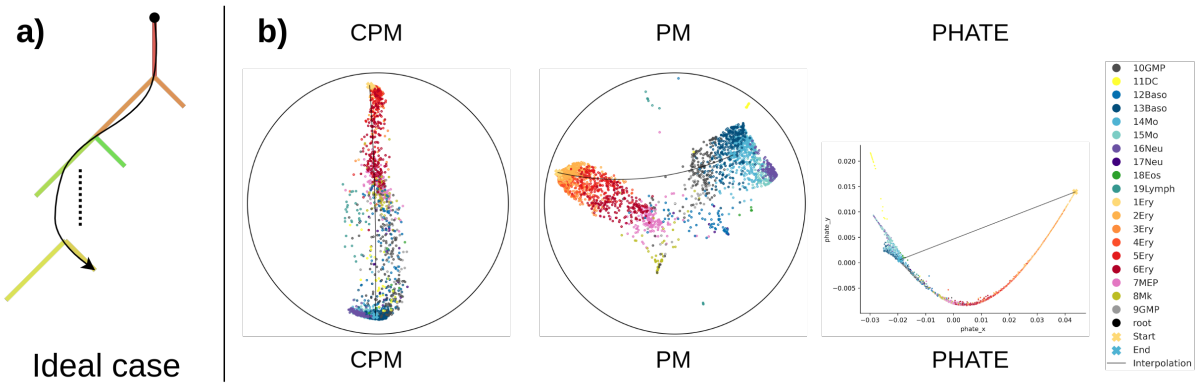

Figure 8: **Interpolation between hierarchical levels.** (a) In hierarchical structures, the path between two nodes should ideally traverse intermediate points that reflect the underlying organization. (b) For the mouse haematopoiesis dataset, only Contrastive Poincaré Maps embedding yields a trajectory that passes through the intermediate Ery and Mk cell types when interpolating between 1Ery and 14Mo.

alizations of embeddings for state-of-the-art methods are shown in Supplementary Figures 17, 23. These results demonstrate the ability of CPM to not only capture intra-lineage hierarchies but also global hierarchies across multiple lineages.

3 Discussion

Robust computational methods are essential for scRNA-seq analysis due to the inherent sparsity, noise, and high dimensionality of single-cell data. Without appropriate models, biological signal is easily confounded by technical variation and dropout effects. As datasets grow in size and complexity, conventional tools lack the scalability and resolution required for accurate cell-type classification, trajectory inference, and gene regulatory analysis. Thus, rigorous computational frameworks are critical to extract meaningful insights and ensure reproducibility across single-cell studies.

Contrastive Poincaré Maps (CPM) is a novel contrastive learning method for learning hyperbolic representation of single-cell data. Hyperbolic representation offers distinct advantages over traditional Euclidean embeddings, particularly for modeling hierarchical structure. Biological systems frequently exhibit nested organization, such as developmental lineages or differentiation trajectories, that Euclidean space struggles to represent without distortion. Hyperbolic geometry, with its exponential volume growth, is inherently suited to capturing such hierarchies with high fidelity. This enables more accurate preservation of both global and local relationships among cells.

Building on these geometric advantages, our method demonstrates several key strengths across synthetic and real scRNA-seq datasets. It achieves competitive representation accuracy without requiring standard pre-processing steps such as PCA, batch correction, or data subsampling, thereby reducing potential sources of bias and information loss. Unlike methods that rely on dimensionality reduction or batch correction (e.g., Scanorama), our approach maintains fidelity to the underlying transcriptomic structure while remaining computationally efficient. It

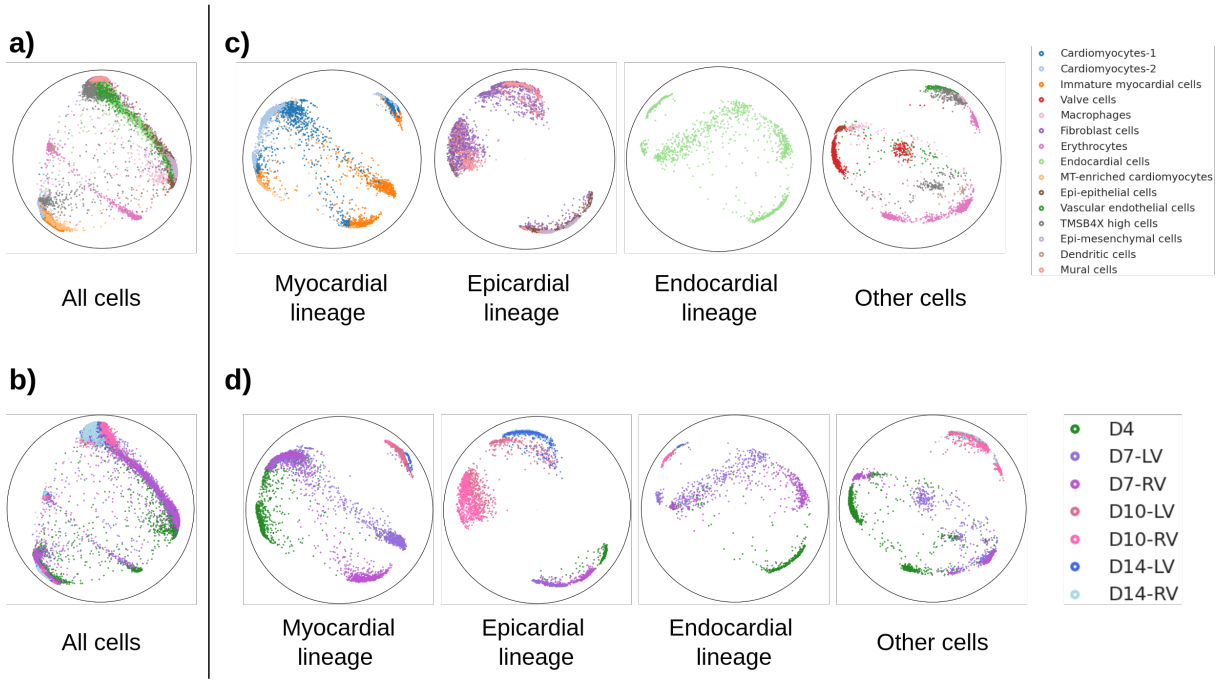

Figure 9: **Chicken cardiogenesis** (a,b) Embedding learned by Contrastive Poincaré Maps for the chicken cardiogenesis dataset, comprising 22,315 cells from multiple lineages across four developmental stages. Contrastive Poincaré Maps preserves the global hierarchical structure of the data. (c,d) Lineage-specific embeddings, highlighting well-preserved intra-lineage hierarchies that capture differentiation trajectories from immature to mature cell types over developmental time.

scales to mouse gastrulation dataset exceeding 100,000 cells Pijuan-Sala et al. (2019) with minimal memory and runtime overhead, enabling robust performance even in large, heterogeneous populations.

Notably, the method accurately captures differentiation trends and complex trajectories in mice hematopoiesis dataset Paul et al. (2015), and multi-lineage hierarchies in chicken cardiogenesis dataset Mantri et al. (2021). This robustness also extends to single-nucleus RNA sequencing data and fine-grained resolution of developmental processes, such as embryonic stage-specific lineage branching during mouse brain development (Supplementary section F). Furthermore, the embedding remains stable across a range of hyperparameter settings (Supplementary Fig. 11), and its geometry enables biologically meaningful interpolation between cell hierarchies, reflecting continuous transitions inherent to cellular differentiation.

While the current framework effectively models hierarchical and continuous structure in scRNA-seq data, several avenues remain open for exploration. A particularly promising direction lies in improving interpretability: unlike widely used visualization tools such as UMAP or t-SNE, whose Euclidean embeddings often suffer from high representation error and distortion of global structure, hyperbolic embeddings preserve both local and global geometry. This preservation opens the door to quantitative interpretability of the latent space, such as assessing biological variation along geodesic paths, offering a principled means to link embedding structure with gene expression dynamics. Future extensions to multimodal data, including spatial transcriptomics and chromatin accessibility, may further enhance the resolution and biological insight afforded by hyperbolic representations.

4 Methods

Let $\mathcal{X} = \{x_i\}_{i=1}^n$ be the given scRNA-seq dataset with n cells and d features per cell, i.e., $x_i \in \mathbb{R}^d$. In this work, we focus on representing data in hyperbolic space by learning a mapping, $\phi : \mathcal{X} \to \mathbb{P}^2$ between data space $\mathcal{X}$ and hyperbolic representation space $\mathbb{P}^2$, which maps a data point $x_i \in \mathcal{X}$ to $\phi(x_i) \in \mathbb{P}^2$.

4.1 Hyperbolic Geometry

Hyperbolic spaces are smooth Riemannian manifolds with constant negative curvature. In this work, we adopt the Poincaré ball and Poincaré half-plane models of hyperbolic space, which are two of several known parametrizations. The *Poincaré ball model* defines a hyperbolic space within the Euclidean unit ball, i.e.,

$$\mathbb{P}^d = \{x \in \mathbb{R}^{d+1} : \|x\| < 1\} \tag{4.1}$$

$$d_{\mathbb{P}}(x, y) = \operatorname{acosh}\left(1 + 2\frac{\|x - y\|^2}{(1 - \|x\|^2)(1 - \|y\|^2)}\right) , \tag{4.2}$$

where $\|\cdot\|$ denotes the usual Euclidean norm. The exponential map $\operatorname{Exp}_x : T_x\mathbb{P}^d \to \mathbb{P}^d$, an important geometric tool, which, e.g., characterises geodesics in the space, is given by

$$\operatorname{Exp}_x(v) = \cosh(\|v\|_M)x + \sinh(\|v\|_M)\frac{v}{\|v\|_M} ,$$

where $\|v\|_M = \sqrt{\langle v, v\rangle_M}$ is computed with respect to the Minkowski product $\langle u, v\rangle_M = -u_0 v_0 + \sum_{i=1}^{d} u_i v_i$. Geodesics in the Poincare model can be computed via simple retractions of the form

$$\begin{aligned} \mathcal{R} : \mathbb{R}^{d+1} &\to T_x \mathbb{P}^d \\ x &\mapsto P(x+v) \end{aligned}$$

where P is an orthogonal projection onto $\mathbb{P}^d$, given by

$$P(v) = \begin{cases} \frac{v}{\|v\|}, \|v\| \leq q \\ v, \text{else} \ . \end{cases}$$

Throughout this work, we will focus on the two-dimensional setting, i.e., $\mathbb{P}^2$.

4.2 Synthetic data generation

We generated synthetic datasets using PHATE Moon et al. (2019), which relies on diffusion-limited aggregation. The ground truth hierarchies were binary trees, with each depth level branching into two subtrees. To introduce heterogeneity, one branch at every level was constrained to contain one-third of the total samples at that depth level. The total number of samples in the is given by n_l(root level) $+(d \times n_l)$, where n_l is the number of samples per level and d is the depth of the tree. We also generated an *imbalanced* tree with randomly sampled branches to simulate more realistic scenarios.

4.3 Data pre-processing

The synthetic data was standardized to have zero mean and unit variance. The mouse gastrulation dataset was normalized to 10000 counts per cell, Log1p transformed and filtered to contain 2000 highly variable genes. The processed mice hematopoiesis dataset was obtained from Klimovskaia et al. Klimovskaia et al. (2020). The chicken cardiogenesis dataset was processed following the approach described in Mantri et al. Mantri et al. (2021), with the exception that CPM operated on the full dataset containing 18,797 genes prior to Scanorama integration and the selection of the top 2000 highly variable genes. While our method operates without the need for additional dimensionality reduction techniques, PCA was applied to the other baselines along with their respective pre-processing steps.

4.4 Model architecture and optimization

Contrastive Poincaré Maps (CPM) adopts sampling strategy of SCARF Bahri et al. (2022). For each batch, a positive sample is obtained by constructing an additional corrupted view: a randomly chosen fraction of feature entities is substituted with values from other training examples. This kind of augmentation strategy contributes to the learning of robust representations across diverse cells. The encoder consists of a projection block and a hyperbolic block. The projection block is composed of linear layers with ReLU activation. The function of the projection block is twofold, acting as both a dimensionality reduction and noise reduction method, which can be adapted and trained end-to-end, unlike conventional methods such as Principal Component Analysis (PCA). The hyperbolic block comprises of hyperbolic linear layers with ReLU activation Chami et al. (2019). We choose a two-dimensional representation space, the Poincaré disk,

with curvature $\kappa = -1$. The choice of a two-dimensional representation space is motivated by ease of visualization and interpretability of the resulting embeddings; varying or learning the curvature as opposed to using a fixed value has been shown to have little effect (see, e.g., Chami et al. (2019)), motivating our choice of $\kappa = -1$. A single forward pass of the model can be written as,

$$H = \sigma(W_e X^e + b_1),$$

where H is the latent representation learnt by the projection block, W_e are the parameters of the projection block, X^e is the input data, b_1 is the bias, and $\sigma(.)$ is the non-linear ReLU activation unit. The representations are subsequently projected to the Poincaré disk using the exponential map $Exp_o(.)$ and processed by the hyperbolic block, as seen in the following equation:

$$Z^h = \sigma^{\otimes}(W_h \otimes \mathrm{Exp}_o(H) \oplus \mathrm{Exp}_o(b_2)) \ .$$

Here, W_h and b_2 are the trainable parameters of the hyperbolic block and Z^h is the hyperbolic representation learnt by the model; $\sigma^{\otimes}(.) = Exp_o(\sigma(Log_o(.)))$, where $Log_o(.)$ is the logarithmic map Chami et al. (2019).

Hyperbolic representations of both augmented views of the input are learnt by using encoder blocks with shared weights. Their similarity, referred to as *Poincaré similarity*, is computed from Poincaré distance (Eq. 4.1) as $1/(1+d_{\mathbb{P}}(x,y))$. These similarities are incorporated into the classical InfoNCE contrastive loss for training. We use Riemannian Adam Bécigneul and Ganea (2023) for optimisation, using different learning rates for the projection and hyperbolic blocks.

4.5 Hyperparameter tuning

The most important hyperparameters include batch size, corruption rate for data augmentation, the number of projection and hyperbolic layers, dimensions of the latent representation and learning rates for the projection and hyperbolic blocks. The learning rate for the hyperbolic block is usually higher than the learning rate for the projection block. The selection of batch size is a trade-off between performance and speed. A batch size of 128 worked well for most of the datasets used in this study. The number of hyperbolic layers depends on the complexity and depth of the input hierarchies. We set the number of hyperbolic layers to 2 for the deepest synthetic tree used in this study. It is also crucial to strike a balance between the number of projection layers and number of cells to be considered for augmentation in a batch, as these control the distortion of our embeddings. A corruption rate of 0.2 and one projection layer achieved good performance on most of the datasets used in this study.

4.6 Complexity

The memory complexity of CPM is $\mathcal{O}(b^2)$, where b is the batch size used for training. The time complexity mainly consists of the calculation of the pairwise similarities for learning the embedding, which is $\mathcal{O}(b^2)$. The time complexity for the entire training process can be summarized as $\mathcal{O}(b^2ed)$, where e is the number of epochs and d is the dimension of feature space. For all the datasets used in this study, the training loss converged within 200 epochs. The training time ranged from a few minutes (< 15) for synthetic datasets, 33-35 minutes for chicken cardiogenesis dataset (22,315 cells and 18,798 genes) to 2-3 hours for mouse gastrulation dataset (>100,000 cells and 2000 genes). The total memory consumption was less than 20GB for all the datasets used in this study.

4.7 Quantitative evaluation metrics

Let $\phi : \mathcal{X} \to \mathcal{M}$ denote an embedding of a data set $\mathcal{X}$. In the case of Poincaré embeddings (standard and contrastive), $\mathcal{M} = \mathbb{P}^2$, i.e., the hyperbolic plane. For Euclidean embeddings, $\mathcal{M} = \mathbb{R}^d$, for small d.

Distortion. Consistent with conventions in the classical metric embedding literature, the representation error of ϕ can be quantified using *distortion*:

$$\frac{1}{c_M} d_{\mathcal{M}}(\phi(x), \phi(y)) \leq d_{\mathcal{X}}(x, y) \leq c_M(\phi) d_{\mathcal{M}}(\phi(x), \phi(y)) \quad \forall x, y \in \mathcal{X} \ .$$

If ϕ is an isometric embedding, then $c_M(\phi) = 1$.

Computational distortion scores. The parameter c_M can be difficult to compute in practice, as access to the map ϕ is required in closed form to evaluate $\phi(x)$ over the whole data space. Therefore, our analysis focuses on the following computational distortion scores (see, e.g., Weber (2020)).

1. **Worst-case distortion:** In certain applications, one may be interested in a worst-case analysis, particularly when robustness guarantees are required across all pairs of points.

$$D_{\text{worst}}(\phi) = \max_{1 \leq i < j \leq |\mathcal{X}|} \left| \left(\frac{d_{\mathcal{M}}(x_i, x_j)}{d_{\mathcal{X}}(x_i, x_j)} \right)^2 - 1 \right| . \tag{4.3}$$

   Here, $d_{\mathcal{M}}$ denotes the distance in the representation space (hyperbolic or Euclidean) and $d_{\mathcal{X}}$ the distance in the data space.

2. **k-MAP score:** The MAP score is a ranking loss, which measures the preservation of local proximity. We build the similarity graph of $\mathcal{X}$ and assess how well the relationship of a point $x \in \mathcal{X}$ with its k-hop neighbors is preserved under the hyperbolic embedding ϕ:

$$\text{MAP}(\phi) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \frac{1}{\deg(x)} \sum_{i=1}^{|\mathcal{N}_k(x)|} \frac{|\mathcal{N}_1(x) \cap R_{x,i}|}{|R_{x,i}|} \in [0, 1] \ . \tag{4.4}$$

   Here, $R_{x,i}$ denotes the smallest set of nearest neighbors required to retrieve the ith neighbor of $\phi(x)$ in $\mathcal{M}$.

If ϕ is an isometric embedding, then $\text{D}_{\text{worst}} = 0$ and $\text{MAP} = 1$.

Q-scores. Q-scores, introduced by Lee et al. Lee and Verleysen (2010), are widely used metrics in the manifold embedding literature. Let $Q = \{q_{kl}\}_{k,l \in [N]}$ be the *co-ranking matrix* defined as

$$\begin{aligned}
q_{kl} &:= |\{(k, l) : \ \mu_{ij} = k, \nu_{ij} = l\}| \\
\mu_{ij} &:= |\{k : \ d_{\mathcal{X}}(x_i, x_k) < d_{\mathcal{X}}(x_i, x_j) \text{ and } \{d_{\mathcal{X}}(x_i, x_k) = d_{\mathcal{X}}(x_i, x_j) \text{ if } 1 \leq k < j \leq N\}\}| \\
\nu_{ij} &:= |\, \{k : \ d_{\mathcal{M}}(x_i, x_k) < d_{\mathcal{M}}(\phi(x_i), \phi(x_j)) \\
&\qquad \text{and } \{d_{\mathcal{M}}(\phi(x_i), \phi(x_k)) = d_{\mathcal{M}}(\phi(x_i), \phi(x_j)) \text{ if } 1 \leq k < j \leq N\}\} \,|
\end{aligned}$$

Here, $|I|$, denotes the cardinality of a set I. The Q-curves are then given by

$$Q(\phi, K) := \frac{1}{KN} \sum_{(k,l)\in I_K} q_{kl} \ , \tag{4.5}$$

which are computed over blocks $I_K := \{1, \ldots, K\} \times \{1, \ldots, K\}$ of the co-ranking matrix, restricted to the upper half of I_K, due to symmetry. The local and global Q-scores of hyperbolic embedding ϕ are computed as

$$K_{\max} := \underset{K}{\operatorname{argmax}} \left(Q(\phi, K) - \frac{K}{N-1} \right) \tag{4.6}$$

$$Q_{local}(\phi) := \frac{1}{K_{\max}} \sum_{K=1}^{K_{\max}} Q(\phi, K) \tag{4.7}$$

$$Q_{global}(\phi) := \frac{1}{N - K_{\max}} \sum_{K=K_{\max}}^{N-1} Q(\phi, K) \ . \tag{4.8}$$

The split of the Q-curve at $K_{\max}$ distinguishes the local from global regimes. Q_{local} and Q_{global} range from 0 (worst) to 1 (best).

4.8 Visualization

Contrastive Poincaré Maps are visualized using the 2D Poincaré disk layout Klimovskaia et al. (2020). To obtain a magnified view, a linear rescaling of the co-ordinates can be applied. By default, the point with the smallest average distance to all others is placed at the origin. Alternatively, the embedding can be translated so that a chosen reference point is mapped to the origin, while maintaining all pairwise hyperbolic distances. Owing to the increased resolution near the origin in hyperbolic space, this translation method is particularly useful for inspecting hierarchical structure when a root is known, or for analyzing cell lineages in multi-lineage datasets. The translation of the entire embedding x to place r at the origin is defined as:

$$\tau_\phi(r, x) = \frac{(1 + 2\langle r, x\rangle + \|x\|^2)r + (1 - \|r\|^2)x}{1 + 2\langle r, x\rangle + \|r\|^2\|x\|^2}$$

4.9 Phylogenetic tree

A simple phylogenetic tree is visualized by plotting low-dimensional Poincaré distances between the cells (as summarized in Equation 4.1) against the embryonic stage. For some of the datasets, individual cells were grouped into lineages. This serves multiple purposes — (i) it helps in understanding the evolutionary trend in different datasets, (ii) it reveals the density of cells at each embryonic stage, (iii) it reveals the diversity within each lineage or cell type.

4.10 Poincaré Half-Plane Model and Geodesic Calculation

We can transition from the Poincaré ball model to the Poincaré half-plane model using an isometry known as the Cayley transform. The Cayley transform preserves the hyperbolic distances

and angles between points, ensuring the integrity of the underlying geometrical structures of the embeddings during the transformation. The Cayley transform is given by

$$C : \mathbb{P}^d \to \mathbb{H}^d, \quad C(x) = \frac{x+i}{ix+1}, \tag{4.9}$$

where $x \in \mathbb{P}^d$ is a point in the Poincaré ball model, i is the imaginary unit, and $\mathbb{H}^d$ is the Poincaré half-plane model.

In the Poincaré half-plane model, geodesics are represented as either vertical lines (when they pass through the point at infinity) or semi-circles that are orthogonal to the boundary line (the x-axis). The geodesic $\gamma : [0,1] \to \mathbb{H}^d$ connecting two points $x, y \in \mathbb{H}^d$ is given by:

$$\gamma(t) = a\cosh(t) + b\sinh(t), \tag{4.10}$$

where a and b are computed to satisfy the boundary conditions $\gamma(0) = x$ and $\gamma(1) = y$. By utilizing these equations, we can compute the geodesic in the half-plane model that connects any two given points.

In a similar manner to our previous experiment, we select pairs of root and leaf nodes and map these points into the half-plane model.

Acknowledgements

The authors would like to thank Alice Bizeul for her input during the initial phase of the project.

Funding: NB is funded by the German Federal Ministry of Health (BMG) within the "Surgomics" project (Grant Number: BMG 2520DAT82) and partially by the German Federal Ministry of Education and Research (BMBF) within the DAAD Konrad Zuse AI school SECAI (School of Embedded Composite AI, https://secai.org/) (Project number: 57616814). MW was funded by NSF awards CBET-2112085 and DMS-2406905, and a Sloan Research Fellowship in Mathematics.

References

Dara Bahri, Heinrich Jiang, Yi Tay, and Donald Metzler. Scarf: Self-supervised contrastive learning using random feature corruption. In *International Conference on Learning Representations*, 2022.

Trygve E Bakken, Rebecca D Hodge, Jeremy A Miller, Zizhen Yao, Thuc Nghi Nguyen, Brian Aevermann, Eliza Barkan, Darren Bertagnolli, Tamara Casper, Nick Dee, Emma Garren, Jeff Goldy, Lucas T. Graybuck, Matthew Kroll, Roger S. Lasken, Kanan Lathia, Sheana Parry, Christine Rimorin, Richard H. Scheuermann, Nicholas J. Schork, Michael Shehata, Soraya I. Tieu, John W. Phillips, Amy Bernard, Kimberly A. Smith, Hongkui Zeng, Ed S. Lein, and Bosiljka Tasic. Single-nucleus and single-cell transcriptomes compared in matched cortical cell types. *PloS one*, 13(12):e0209648, 2018.

Gary Bécigneul and Octavian-Eugen Ganea. Riemannian adaptive optimization methods. In *International Conference on Learning Representations (ICLR 2019)*, volume 9, pages 6384–6399. Curran, 2023.

Volker Bergen, Marius Lange, Stefan Peidli, F Alexander Wolf, and Fabian J Theis. Generalizing rna velocity to transient cell states through dynamical modeling. *Nature biotechnology*, 38(12): 1408–1414, 2020.

Nithya Bhasker, Hattie Chung, Louis Boucherie, Vladislav Kim, Stefanie Speidel, and Melanie Weber. Contrastive poincaré maps for single-cell data analysis. In *ICLR 2024 Workshop on Machine Learning for Genomics Explorations*, 2024.

Benjamin Paul Chamberlain, James Clough, and Marc Peter Deisenroth. Neural embeddings of graphs in hyperbolic space. *arXiv preprint arXiv:1705.10359*, 2017.

Ines Chami, Zhitao Ying, Christopher Ré, and Jure Leskovec. Hyperbolic graph convolutional neural networks. *Advances in neural information processing systems*, 32, 2019.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PmLR, 2020.

R. R. Coifman, S. Lafon, A. B. Lee, M. Maggioni, B. Nadler, F. Warner, and S. W. Zucker. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences*, 102(21):7426–7431, May 2005.

Joakim S Dahlin, Fiona K Hamey, Blanca Pijuan-Sala, Mairi Shepherd, Winnie WY Lau, Sonia Nestorowa, Caleb Weinreb, Samuel Wolock, Rebecca Hannah, Evangelia Diamanti, David G. Kent, Berthold Gottgens, and Nicola K. Wilson. A single-cell hematopoietic landscape resolves 8 lineage trajectories and defects in kit mutant mice. *Blood, The Journal of the American Society of Hematology*, 131(21):e1–e11, 2018.

Octavian Ganea, Gary Bécigneul, and Thomas Hofmann. Hyperbolic neural networks. *Advances in neural information processing systems*, 31, 2018.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, 2021.

Aritra Ghosh and Andrew Lan. Contrastive learning improves model robustness under label noise. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2703–2708, 2021.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Azar Gheshlaghi, Bilal Piot, Koray Kavukcuoglu, Rémi Munso, and Michal Valko. Bootstrap your own latent-a new approach to self-supervised learning. *Advances in neural information processing systems*, 33:21271–21284, 2020.

Rashel V Grindberg, Joyclyn L Yee-Greenbaum, Michael J McConnell, Mark Novotny, Andy L O'Shaughnessy, Georgina M Lambert, Marcos J Araúzo-Bravo, Jun Lee, Max Fishman, Gillian E Robbins, Xiaoying Lin, Pratap Venepally, Jonathan H. Badger, David W. Galbraith, Fred H. Gage, and Roger S. Lasken. Rna-sequencing from single nuclei. *Proceedings of the National Academy of Sciences*, 110(49):19802–19807, 2013.

Anupam Gupta. Embedding tree metrics into low dimensional euclidean spaces. In *Proceedings of the Thirty-First Annual ACM Symposium on Theory of Computing*, STOC '99, page 694–700, New York, NY, USA, 1999. Association for Computing Machinery. ISBN 1581130678.

Naomi Habib, Inbal Avraham-Davidi, Amit Basu, Tyler Burks, Karthik Shekhar, Matan Hofree, Suvrit Ray Choudhury, Francois Aguet, Eran Gelfand, Kristin Ardlie, David A. Weitz, Orit Rozenblatt-Rosen, Feng Zhang, and Aviv Regev. Massively parallel single-nucleus rna-seq with dronc-seq. *Nature Methods*, 14(10):955–958, 2017.

Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

Laleh Haghverdi, Maren Büttner, F. Alexander Wolf, Florian Buettner, and Fabian J. Theis. Diffusion pseudotime robustly reconstructs lineage branching. *Nature Methods*, 13(10):845–848, 2016.

Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9729–9738, 2020.

Brian L Hie, Soochi Kim, Thomas A Rando, Bryan Bryson, and Bonnie Berger. Scanorama: integrating large and diverse single-cell transcriptomic datasets. *Nature protocols*, 19(8):2283–2297, 2024.

Peter V. Kharchenko. The triumphs and limitations of computational methods for scRNA-seq. *Nature Methods*, 18(7):723–732, 2021.

Anna Klimovskaia, David Lopez-Paz, Léon Bottou, and Maximilian Nickel. Poincaré maps for analyzing complex hierarchies in single-cell data. *Nature communications*, 11(1):2966, 2020.

Aleksandra A Kolodziejczyk, Jong Kyoung Kim, Valentine Svensson, John C Marioni, and Sarah A Teichmann. The technology and biology of single-cell rna sequencing. *Molecular Cell*, 58(4):610–620, 2015.

Joseph B. Kruskal and Myron Wish. *Multidimensional Scaling*. SAGE, 1978.

Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E. Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Kharchenko Peter V. Rna velocity of single cells. *Nature*, 560(7719):494–498, 2018a.

Gioele La Manno, Ruslan Soldatov, Amit Zeisel, Emelie Braun, Hannah Hochgerner, Viktor Petukhov, Katja Lidschreiber, Maria E Kastriti, Peter Lönnerberg, Alessandro Furlan, Jean Fan, Lars E. Borm, Zehua Liu, David van Bruggen, Jimin Guo, Xiaoling He, Roger Barker, Erik Sundström, Gonçalo Castelo-Branco, Patrick Cramer, Igor Adameyko, Sten Linnarsson, and Kharchenko Peter V. Rna velocity of single cells. *Nature*, 560(7719):494–498, 2018b.

Blue B Lake, Rizi Ai, Gwendolyn E Kaeser, Neeraj S Salathia, Yun C Yung, Rui Liu, Andre Wildberg, Derek Gao, Ho-Lim Fung, Song Chen, Raakhee Vijayaraghavan, Julian Wong, Allison Chen, Xiaoyan Sheng, Fiona Kaper, Richard Shen, Mostafa Ronaghi, Jian-Bing Fan, Wei Wang, Jerold Chun, and Kun Zhang. Neuronal subtypes and diversity revealed by single-nucleus rna sequencing of the human brain. *Science*, 352(6293):1586–1590, 2016.

John A. Lee and Michel Verleysen. Scale-independent quality criteria for dimensionality reduction. *Pattern Recognition Letters*, 31(14):2248–2257, October 2010.

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

Evan Z Macosko, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, Allison R Bialas, Nolan Kamitaki, Emily M Martersteck, John J Trombetta, David A Weitz, Joshua R Sanes, Alex K Shalek, Aviv Regev, and Steven A McCaroll. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell*, 161 (5):1202–1214, 2015.

Madhav Mantri, Gaetano J Scuderi, Roozbeh Abedini-Nassab, Michael FZ Wang, David McKellar, Hao Shi, Benjamin Grodner, Jonathan T Butcher, and Iwijn De Vlaminck. Spatiotemporal single-cell rna sequencing of developing chicken hearts identifies interplay between cellular differentiation and morphogenesis. *Nature communications*, 12(1):1771, 2021.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861, 2018.

Kevin R Moon, David van Dijk, Zheng Wang, Scott Gigante, Daniel B Burkhardt, William S Chen, Kristina Yim, Antonia van den Elzen, Matthew J Hirn, Ronald R Coifman, Natalia B. Ivanova, Guy Wolf, and Smita Krishnaswamy. Visualizing structure and transitions in high-dimensional biological data. *Nature biotechnology*, 37(12):1482–1492, 2019.

Samantha A Morris. The evolving concept of cell identity in the single cell era. *Development*, 146(12):dev169748, 2019.

Boaz Nadler, Stephane Lafon, Ioannis Kevrekidis, and Ronald Coifman. Diffusion Maps, Spectral Clustering and Eigenfunctions of Fokker-Planck Operators. In *Advances in Neural Information Processing Systems*, volume 18. MIT Press, 2005.

Sonia Nestorowa, Fiona K Hamey, Blanca Pijuan Sala, Evangelia Diamanti, Mairi Shepherd, Elisa Laurenti, Nicola K Wilson, David G Kent, and Berthold Göttgens. A single-cell resolution map of mouse hematopoietic stem and progenitor cell differentiation. *Blood, The Journal of the American Society of Hematology*, 128(8):e20–e31, 2016.

Maximillian Nickel and Douwe Kiela. Poincaré embeddings for learning hierarchical representations. *Advances in neural information processing systems*, 30, 2017.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*, 2018.

Franziska Paul, Ya'ara Arkin, Amir Giladi, Diego Adhemar Jaitin, Ephraim Kenigsberg, Hadas Keren-Shaul, Deborah Winter, David Lara-Astiaso, Meital Gury, Assaf Weiner, Eyal David, Nadav Cohen, Felicia Kathrine Bratt Lauridsen, Simon Haas, Andreas Schlitzer, Alexander Mildner, Florent Ginhoux, Steffen Jung, Andreas Trumpp, Bo Torben Porse, Amos Tanay, and Ido Amit. Transcriptional heterogeneity and lineage commitment in myeloid progenitors. *Cell*, 163(7):1663–1677, 2015.

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

Blanca Pijuan-Sala, Jonathan A Griffiths, Carolina Guibentif, Tom W Hiscock, Wajid Jawaid, Fernando J Calero-Nieto, Carla Mulas, Ximena Ibarra-Soria, Richard C V Tyser, Debbie Lee Lian Ho, Wolf Reik, Shankar Srinivas, Benjamin D Simons, Jennifer Nichols, John C Marioni, and Berthold Göttgens. A single-cell molecular map of mouse gastrulation and early organogenesis. *Nature*, 566(7745):490–495, 2019.

John J Quinn, Megan G Jones, Rikiya A Okimoto, Shigeki Nanjo, Mindy M Chan, Nir Yosef, Trever G Bivona, and Jonathan S Weissman. Single-cell lineages reveal the rates, routes, and drivers of metastasis in cancer xenografts. *Science*, 371(6532):eabc1944, 2021.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

Markus Ringnér. What is principal component analysis? *Nature Biotechnology*, 26(3):303–304, 2008.

Andrew JC Russell, Jackson A Weir, Naeem M Nadaf, Matthew Shabet, Vipin Kumar, Sandeep Kambhampati, Ruth Raichur, Giovanni J Marrero, Sophia Liu, Karol S Balderrama,

Charles R. Vanderburg, Vignesh Shanmugam, Luyi Tian, J. Bryan Iorgulescu, Charles H. Yoon, Catherine J. Wu, Evan Z. Macosko, and Fei Chen. Slide-tags enables single-nucleus barcoding for multimodal spatial genomics. *Nature*, 625(7993):101–109, 2024.

Wouter Saelens, Robrecht Cannoodt, Helena Todorov, and Yvan Saeys. A comparison of single-cell trajectory inference methods. *Nature biotechnology*, 37(5):547–554, 2019.

Frederic Sala, Chris De Sa, Albert Gu, and Christopher Ré. Representation tradeoffs for hyperbolic embeddings. In *International conference on machine learning*, pages 4460–4469. PMLR, 2018.

Rik Sarkar. Low distortion delaunay embedding of trees in hyperbolic plane. In *International Conference on Graph Drawing*, page 355–366, 2011.

Fuchou Tang, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, John Bodeau, Brian B Tuch, Asim Siddiqui, Kaiqin Lao, and M Azim Surani. mrna-seq whole-transcriptome analysis of a single cell. *Nature methods*, 6(5):377–382, 2009.

Cole Trapnell. Defining cell types and states with single-cell genomics. *Genome research*, 25 (10):1491–1498, 2015.

Cole Trapnell, Davide Cacchiarelli, Jonna Grimsby, Prapti Pokharel, Shuqiang Li, Michael Morse, Niall J Lennon, Kenneth J Livak, Tarjei S Mikkelsen, and John L Rinn. The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells. *Nature biotechnology*, 32(4):381–386, 2014.

Barbara Treutlein, Doug G Brownfield, Angela R Wu, Norma F Neff, Gary L Mantalas, F Hernan Espinoza, Tushar J Desai, Mark A Krasnow, and Stephen R Quake. Reconstructing lineage hierarchies of the distal lung epithelium using single-cell rna-seq. *Nature*, 509(7500):371–375, 2014.

Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

Melanie Weber. Neighborhood growth determines geometric priors for relational representation learning. In *International Conference on Artificial Intelligence and Statistics*, pages 266–276. PMLR, 2020.

Melanie Weber, Manzil Zaheer, Ankit Singh Rawat, Aditya K Menon, and Sanjiv Kumar. Robust large-margin learning in hyperbolic space. *Advances in Neural Information Processing Systems*, 33:17863–17873, 2020.

F Alexander Wolf, Fiona K Hamey, Mireya Plass, Jordi Solana, Joakim S Dahlin, Berthold Göttgens, Nikolaus Rajewsky, Lukas Simon, and Fabian J Theis. Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. *Genome biology*, 20:1–9, 2019.

A Background

A.1 Single-cell RNA sequencing

Single-cell RNA sequencing (scRNA-seq) is a transformative technique that allows for transcriptomic profiling at the resolution of individual cells, enabling researchers to characterize cellular heterogeneity, identify rare cell populations, and reconstruct developmental trajectories Tang et al. (2009); Kolodziejczyk et al. (2015). Unlike bulk RNA-seq, which averages gene expression across many cells, scRNA-seq captures cell-specific transcriptional signatures, crucial for understanding complex biological systems, especially during dynamic processes such as development and differentiation Trapnell (2015).

In multicellular organisms, cell differentiation is the process through which progenitor or stem cells acquire specialized functions and identities. This is a tightly regulated process involving coordinated changes in gene expression, often orchestrated by lineage-specific transcription factors, epigenetic modifications, and signaling pathways Morris (2019). scRNA-seq offers a unique opportunity to capture these dynamic changes at high resolution, enabling the identification of intermediate cell states and transient gene expression programs that are otherwise obscured in population-level analyses La Manno et al. (2018b).

One of the most critical applications of scRNA-seq in the context of cell differentiation is lineage detection, which is the inference or experimental tracing of the developmental history of individual cells. Accurate lineage reconstruction provides insight into the hierarchy of cell fate decisions and the molecular determinants that guide them. Computational approaches such as pseudotime analysis Trapnell et al. (2014), trajectory inference Saelens et al. (2019), and RNA velocity Bergen et al. (2020) have been developed to infer the progression of differentiation from transcriptomic data. These methods attempt to order cells along a continuum reflecting a temporal or developmental progression, based on similarities in gene expression patterns.

Understanding lineage relationships is particularly vital in developmental systems biology, regenerative medicine, and cancer research, where the misregulation of differentiation programs can lead to developmental disorders. For example, in hematopoiesis, scRNA-seq has revised the classical view of hierarchical differentiation into a more continuous landscape of lineage priming and bifurcations Paul et al. (2015). Similarly, in cancer, lineage tracing with scRNA-seq has elucidated how tumors recapitulate developmental programs and acquire resistance through plastic differentiation Quinn et al. (2021).

In summary, scRNA-seq has become an indispensable tool for dissecting the cellular basis of differentiation. Lineage detection is essential for interpreting scRNA-seq data in a developmental context and for uncovering the principles that govern cell fate decisions.

A.2 Single nucleus RNA sequencing

Single nucleus RNA sequencing (snRNA-seq) has emerged as a powerful alternative to single-cell RNA sequencing (scRNA-seq) for transcriptomic profiling of individual cells, particularly in tissues that are difficult to dissociate or where intact cell isolation is impractical Habib et al. (2017). Unlike scRNA-seq, which relies on whole-cell dissociation, snRNA-seq isolates and profiles RNA from individual nuclei, thereby minimizing stress-induced transcriptional artifacts and enabling the study of archived or frozen specimens Bakken et al. (2018); Lake et al. (2016). This approach preserves the native transcriptional landscape of cells from complex or fragile

tissues such as the adult brain, fibrotic tissue, and solid tumors Grindberg et al. (2013). These advantages make snRNA-seq an indispensable tool for characterizing cellular heterogeneity in contexts where conventional scRNA-seq is limited.

A.3 Euclidean Embedding Methods

Analyzing complex biological datasets derived from next-generation sequencing technologies is often challenging, given their high-dimensional nature. To make exploratory analysis more tractable, dimension reduction techniques are commonly used to generate a low-dimensional representation of the data (Kharchenko (2021)) that approximates pairwise distances in the original high-dimensional space. For instance, multidimensional scaling (MDS) Kruskal and Wish (1978) projects the original data $X \in \mathbb{R}^{n \times m}$ with n observations and m dimensions onto a low-dimensional space $Y \in \mathbb{R}^{n \times p}, p << m$ so that pairwise Euclidean distances in the embedding space $\|y_i - y_j\|_{\mathbb{E}}$ approximate the distances $d_{ij} = \|x_i - x_j\|$ in the input space:

$$\underset{y_1,\ldots,y_n}{\arg\min} \sum_{i<j} (\|y_i - y_j\|_{\mathbb{E}} - d_{ij})^2.$$

Principal component analysis (PCA) Ringnér (2008) is a special case of MDS with the Euclidean metric used in the input space. In practice, PCA is computed via the eigendecomposition of the covariance matrix $\mathrm{cov}(X) = \frac{1}{n-1} X^{\mathrm{T}} X = V \Sigma V^{\mathrm{T}}$. By applying PCA, the data is linearly transformed $T = XV_p$ with $V_p \in \mathbb{R}^{m \times p}$, resulting in the alignment of the first p principal components with the axes of maximal variation.

While PCA reveals the global structure of the data, its linearity assumption makes it inadequate to accurately represent nonlinear low-dimensional structure in complex biological data sets. As a result, nonlinear dimension reduction techniques t-SNE Maaten and Hinton (2008)) and UMAP McInnes et al. (2018) have garnered popularity in single-cell transcriptome analyses. Unlike PCA, t-SNE and UMAP favor the preservation of local distances at the cost of global distances. To achieve this, t-SNE approximates Gaussian kernel similarities p_{ij} in the input space using Student's t-distribution similarities q_{ij} in the embedding space. Similarly, UMAP constructs a k-nearest neighbor (KNN) graph in the input space and approximates the manifold spanned by this graph using the low-dimensional map. Both t-SNE and UMAP rely on Euclidean distances $\|y_i - y_j\|_{\mathbb{E}}$ in the embedding space to determine pairwise similarities and fuzzy set memberships, respectively. Thus, the Euclidean intuition indirectly influences the geometry of t-SNE and UMAP embeddings.

Another class of algorithms extensively used to visualize single-cell data is based on the diffusion process Coifman et al. (2005); Nadler et al. (2005), representing a random walk along the KNN graph, which models the underlying data manifold. Within this framework, the edges in the KNN graph are assigned weights corresponding to pairwise similarities p_{ij}, allowing the construction of a transition matrix P, the diffusion operator, for the random walk process. The conventional diffusion map emerges through the eigendecomposition of this transition matrix P Coifman et al. (2005). Advancements in this research field led to the development of diffusion pseudotime (DPT) Haghverdi et al. (2016)) and PHATE Moon et al. (2019) algorithms, both of which are widely used for single-cell lineage tracing. However, even in sophisticated methods such as PHATE, the reliance on Euclidean distances in the embedding space persists. Despite their advantages and well-established methodologies, Euclidean embedding methods may

face challenges in effectively capturing hierarchical structures, often present in biological data. Leveraging non-Euclidean geometries can address some of these challenges.

A.4 Representing Data in Hyperbolic Space

A.4.1 Hyperbolic Embeddings

In this work, we focus on representing data in two-dimensional hyperbolic space, by learning a map $\phi : \mathcal{X} \to \mathbb{P}^2$ between data space $\mathcal{X}$ and hyperbolic representation space $\mathbb{P}^2$, which maps a data point $x \in \mathcal{X}$ to $\phi(x) \in \mathbb{P}^2$. Classical mathematics literature provides theoretical guarantees on the accuracy of representing trees in low-dimensional Euclidean or hyperbolic space. Here, representation error refers to the extent to which pairwise distances among the data points are preserved between the data space and the representation space. Gupta et al. Gupta (1999) demonstrated that a tree with L leaves (i.e., number of nodes in the last level of the tree) cannot be represented in d-dimensional Euclidean space with representation error better than a multiplicative factor of $O(L^{1/(d-1)})$. In contrast, Sarkar et al. Sarkar (2011) showed that trees of *any* size can be embedded into the two-dimensional hyperbolic space with near-exact distance preservation, i.e., with representation error no worse than a multiplicative factor of $O(1+\epsilon)$.

Hierarchical or tree-structured data are common in machine learning, and the complexity of algorithms often scales with the dimension of the data space. Hyperbolic geometry is particularly appealing because it allows for faithful representation of hierarchical data in low dimensions, making it promising for efficient data analysis. A growing line of work explore methods for computing hyperbolic representations of data Nickel and Kiela (2017); Chamberlain et al. (2017); Chami et al. (2019). Throughout this work, we focus on the two-dimensional Poincaré disk setting, $\mathbb{P}^2$. In some cases, we will represent data using a second parametrization, the closely related *Poincaré half-plane model*. In two dimensions, it is given by

$$\mathbb{H}^2 = \{x \in \mathbb{R}^2 \,:\, x_1 > 0\}$$
$$d_{\mathbb{H}}(x, x') = \operatorname{acosh}\left(1 + \frac{(x_0' - x_0)^2 + (x_1' - x_1)^2}{2x_1 x_1'}\right) .$$

Importantly, the two parameterizations are isometric, i.e., we can map point configurations from $\mathbb{P}^2$ to $\mathbb{H}^2$ distortion-free.

A.4.2 Geodesic interpolation

In order to investigate the intrinsic structure of hyperbolic embeddings, it can be beneficial to interpolate between points along geodesics, which constitute the shortest paths in this space. In the hyperbolic geometry of the Poincaré ball model, geodesics are arcs of circles that intersect the boundary of the ball orthogonally. For any two points in the Poincaré ball, there exists a unique geodesic connecting them. To validate the utility of our hyperbolic embeddings, we conduct an experiment involving geodesic interpolation. Specifically, we select pairs of root and leaf nodes in our hierarchical tree-like data, and compute the unique geodesic path that connects these two points in the hyperbolic space.

This interpolation experiment allows us to examine the relationship between data points and the underlying structure of the embedded data. By considering the root and leaf nodes, we are effectively exploring the extremes of the hierarchical structure encoded by our hyperbolic

embeddings. The intermediate points along the geodesic offer valuable insights into the transition from the root to the leaf, and potentially reflect the tree's structure (Fig. 8).

For each pair of root and leaf nodes, we perform the following steps:

1. Compute the coordinates of the root and leaf nodes in the hyperbolic embedding space.
2. Compute the unique geodesic $\gamma(t)$ that connects the root to the leaf.
3. Select a set of t values to sample intermediate points along the geodesic.
4. For each intermediate point, compute its hyperbolic distance to the geodesic.

We expect that if the hyperbolic embeddings have captured the tree structure well, then the distances from the intermediate points to the geodesic should be relatively small, reflecting a high consistency of the tree structure with the geodesic path in the embedding space. On the other hand, if these distances are large, it would suggest that the embedding may not fully capture the hierarchical relationships in the original data. The results from this experiment provide us with a quantitative measure of the quality of our embeddings and the effectiveness of hyperbolic space in capturing the hierarchical structure of tree-like data.

The computation of geodesics is as follows: Given two points $x, y \in \mathbb{P}^d$, the geodesic $\gamma : [0, 1] \to \mathbb{P}^d$ from x to y is formulated by

$$\gamma(t) = \mathrm{Exp}_x(t \cdot \mathrm{Log}_x(y)) \ , \tag{A.1}$$

where $\mathrm{Exp}_x(v)$ and $\mathrm{Log}_x(y)$ refer to the exponential and logarithm maps at x, respectively. Both maps can be expressed in terms of hyperbolic cosine and sine functions and the Minkowski product. In particular,

$$\mathrm{Exp}_x(v) = \cosh(\|v\|_M)x + \sinh(\|v\|_M)\frac{v}{\|v\|_M} \ , \tag{A.2}$$

and

$$\mathrm{Log}_x(y) = \frac{d_{\mathbb{P}}(x, y)}{\sinh(d_{\mathbb{P}}(x, y))}(y - \cosh(d_{\mathbb{P}}(x, y))x) \ . \tag{A.3}$$

These equations allow for the calculation of a point along the geodesic from x to y at any parameter $t \in [0, 1]$. For $t = 0$, $\gamma(t)$ returns x; for $t = 1$, it returns y. For values of t in between, it yields a point along the geodesic from x to y.

Geodesic interpolation thus provides a meaningful mechanism to traverse the learnt representation space, enabling a better understanding and visualization of the structure and relationship of the embedded data in the hyperbolic space.

A.5 Contrastive Learning

Contrastive learning is a self-supervised learning paradigm that aims to learn an embedding space in which similar data pairs (positive samples) are mapped to nearby points and dissimilar pairs (negative samples) are mapped far apart. It is grounded in the idea of learning by comparison, drawing from concepts in metric learning and information theory Hadsell et al. (2006); Oord et al. (2018).

Let $\mathcal{X}$ denote the input space and $f_\theta : \mathcal{X} \to \mathbb{R}^d$ be an encoder parameterized by θ that maps inputs to a d-dimensional representation space. For a given anchor sample $x_i \in \mathcal{X}$, a positive sample x_i^+ (e.g., an augmented version of x_i) and a set of negative samples $\{x_j^-\}_{j=1}^{N-1}$ are used to define the contrastive objective.

A common objective function is the InfoNCE loss Oord et al. (2018), given by:

$$\mathcal{L}_i = -\log \frac{\exp\left(\mathrm{sim}(f_\theta(x_i), f_\theta(x_i^+))/\tau\right)}{\sum_{j=1}^{N} \exp\left(\mathrm{sim}(f_\theta(x_i), f_\theta(x_j^-))/\tau\right)}$$

where $\mathrm{sim}(u, v) = \frac{u^\top v}{\|u\|\|v\|}$ is the cosine similarity, and $\tau > 0$ is a temperature parameter.

Contrastive learning has demonstrated strong performance in various domains. Unsupervised pre-training of image encoders such as SimCLR Chen et al. (2020), MoCo He et al. (2020) provides competitive performance on downstream tasks such as image classification, object detection, and segmentation, compared to supervised baselines. In Natural Language Processing, sentence embeddings Gao et al. (2021) and document retrieval benefit from contrastive objectives using positive pairs from natural augmentations or dropout noise. In general, contrastive learning has emerged as a key technique for effective representation learning in the absence of labels, enabling models to leverage large unlabeled datasets and generalize well to downstream supervised tasks.

SCARF The lack of suitable and meaningful augmentation strategies is a key limitation for applying contrastive learning to tabular data. SCARF Bahri et al. (2022) introduces a contrastive learning framework specifically designed for tabular data, addressing this key limitation in the domain. While contrastive learning has achieved remarkable success in modalities like vision and language through domain-specific augmentations like image cropping or token masking, such techniques do not transfer directly to tabular formats. SCARF tackles this by proposing a novel data augmentation method based on random feature corruption. For each data instance, a positive pair is generated by randomly replacing a subset of its features with values sampled from the empirical marginal distributions of those features. This ensures the augmented sample remains in-distribution while introducing variability necessary for contrastive training.

This corruption-based augmentation is domain-agnostic, making it highly applicable across diverse tabular datasets without requiring handcrafted feature engineering. The learning objective encourages the model to generate invariant representations between the original and corrupted views, enabling robust representation learning. Notably, this approach is well-suited to challenges typical in tabular data, such as noise, missing values, and high feature heterogeneity.

B Baselines

t-SNE T-distributed stochastic neighbor embedding (t-SNE) is a non-linear dimensionality reduction method Van der Maaten and Hinton (2008) that approximates Gaussian kernel similarities p_{ij} in the input space using heavier-tailed Student's t-distribution similarities q_{ij} in the lower-dimensional map:

$$p_{ij} = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$$

$$q_{ij} = \frac{(1 + \|\mathbf{y}_i - \mathbf{y}_j\|^2)^{-1}}{\sum_{k \neq l}(1 + \|\mathbf{y}_k - \mathbf{y}_l\|^2)^{-1}}$$

Perplexity is the key parameter in t-SNE that determines the dispersion of the Gaussian kernel and indirectly influences the balance between preserving the local and global structure of the data. We used the Scikit-learn v1.0.2 Pedregosa et al. (2011) implementation of t-SNE with the Barnes-Hut approximation and varied perplexity as the main hyperparameter. The perplexity values tested were $[5, 10, 20, 30, 40, 50]$.

UMAP Uniform manifold approximation and projection (UMAP) is a non-linear dimensionality reduction method that combines manifold learning and topological data analysis techniques McInnes et al. (2018). It constructs a fuzzy topological representation by transforming the high-dimensional data into a k-nearest neighbor (KNN) graph with edges weighted by the normalized Gaussian kernel similarities:

$$p_{ij} = \frac{\exp\left(-\frac{\|x_i - x_j\|^2}{\sigma_i}\right)}{\sum_{k \neq i} \exp\left(-\frac{\|x_i - x_k\|^2}{\sigma_i}\right)}$$

The low-dimensional embedding in UMAP is obtained by minimizing the cross-entropy between the high-dimensional (p_{ij}) and low-dimensional (q_{ij}) fuzzy topological representations:

$$C = -\sum_{i \neq j} p_{ij} \log(q_{ij}) + (1 - p_{ij}) \log(1 - q_{ij})$$

The key hyperparameters of UMAP are the number of nearest neighbors k and minimum distance d_{min} between the points in the low-dimensional embedding space. We used the python implementation UMAP v0.5.3 McInnes et al. (2018).

PHATE PHATE is a diffusion-based method commonly used for single-cell trajectory analysis Moon et al. (2019). Similar to UMAP, PHATE constructs a KNN graph with edges weighted by normalized kernel similarities $p_{ij} = K_{ij} / \sum_{k \neq i} K_{ik}$. Unlike UMAP, PHATE uses an alpha-decaying Gaussian kernel:

$$K_{ij} = \frac{1}{2} \exp\left(-\left(\frac{||x_i - x_j||}{\sigma_i}\right)^{\alpha}\right) + \frac{1}{2} \exp\left(-\left(\frac{||x_i - x_j||}{\sigma_j}\right)^{\alpha}\right)$$

PHATE differs from the classical diffusion maps by using the diffusion potential $D_{ij}(t) = [P^t]_{ij}$ instead of the diffusion operator P itself. The exponent t, automatically chosen by the algorithm, represents the number of steps taken in a random walk and the powers of the diffusion operator P^t capture structure of the data at different scales. To obtain a low-dimensional embedding Y, PHATE applies multidimensional scaling (MDS) to the diffusion potential D.

Poincare Maps Klimovskaia et al. Klimovskaia et al. (2020) propose Poincaré maps, a dimensionality reduction technique that embeds single-cell RNA sequencing (scRNA-seq) data into a hyperbolic space to better capture hierarchical relationships. Unlike conventional methods such as t-SNE or UMAP that operate in Euclidean spaces and often struggle with representing branching or hierarchical structures, Poincaré maps utilize the Poincaré disk model of hyperbolic geometry to preserve both local and global topological features.

The method begins by constructing a k-nearest neighbor graph over d-dimensional expression data $\mathcal{X} = \{\mathbf{x}_i\}_{i=1}^n$, where $\mathbf{x}_i \in \mathbb{R}^d$. Geodesic distances are then approximated via shortest paths within this graph to reflect the manifold structure of the data. These distances are used to guide the embedding of the data into the two-dimensional Poincaré disk $\mathbb{D}^2$ by minimizing the Kullback–Leibler divergence between the geodesic distances in the input space and the hyperbolic distances in the embedded space.

Empirical results demonstrate that Poincaré maps outperform standard techniques in representing branching structures in various scRNA-seq datasets. However, Poincaré maps are a shallow embedding method, meaning that the embedding is computed directly from pairwise distance relationships without the use of deep neural networks or hierarchical feature extraction layers. This inhibits the model to capture complex nonlinear transformations compared to deep embedding methods.

C Additional results: Model robustness

We evaluated the performance of all embedding methods across all datasets without applying PCA-based dimensionality reduction (Supplementary Fig. 10). For the chicken cardiogenesis dataset, batch correction with Scanorama Hie et al. (2024) was omitted. Performance declined for all methods except CPM, which remained robust in the absence of additional pre-processing.

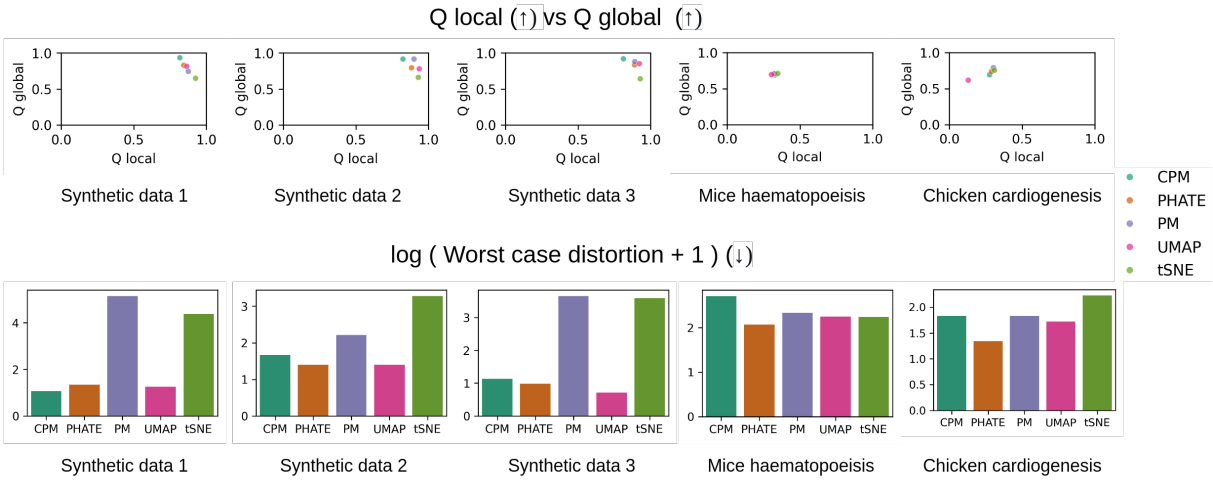

Figure 10: **Robustness to pre-processing** Performance of all embedding methods evaluated without PCA-based dimensionality reduction across all datasets. For the chicken cardiogenesis dataset, batch correction using Scanorama was also omitted.

We also evaluated the performance of Contrastive Poincaré Maps under varying model hyperparameters (Supplementary Fig. 11). The quality metrics remain largely consistent across different configurations, demonstrating the robustness of the model to hyperparameter choices.

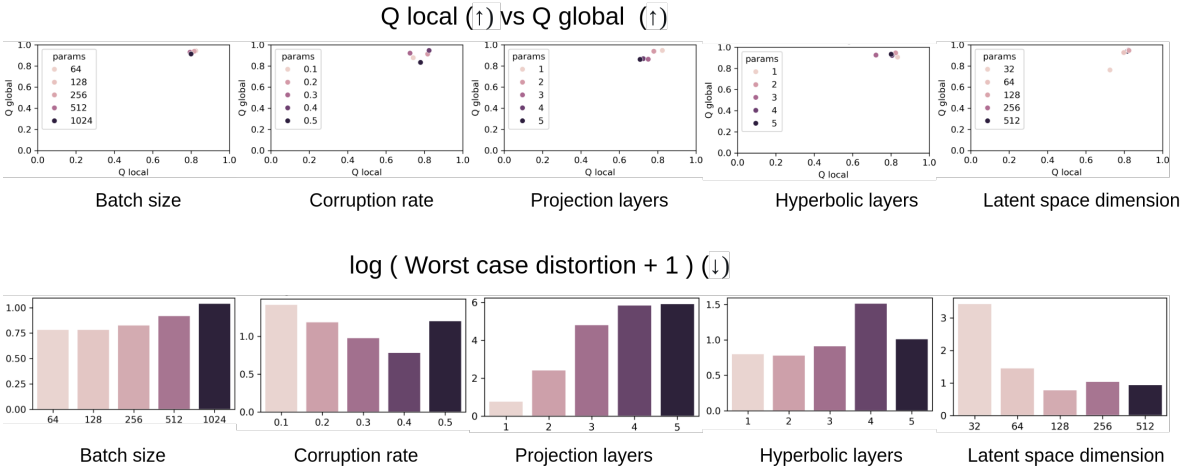

Figure 11: **Robustness of Contrastive Poincaré Maps to hyperparameter variation** Performance of Contrastive Poincaré Maps across a range of model hyperparameter configurations.

D Additional results: Visualizations of embeddings

This section presents visualizations of embeddings generated by all methods across all datasets. For synthetic datasets, visualizations are colour-coded by branch, with the corresponding ground truth tree structure included for reference. For real datasets, embeddings are colour-coded by cell type to reflect known biological annotations. There is no continuous progression between hierarchical levels in Poincaré maps, as demonstrated in CPM. All hierarchical levels beyond the root appear to be pushed toward the horizon, resulting in an unnatural representation that fails to preserve the inherent hierarchical structure. On closer inspection, we also see an unnatural intersection between branches on some of the visualizations produced by PHATE, t-SNE and UMAP.

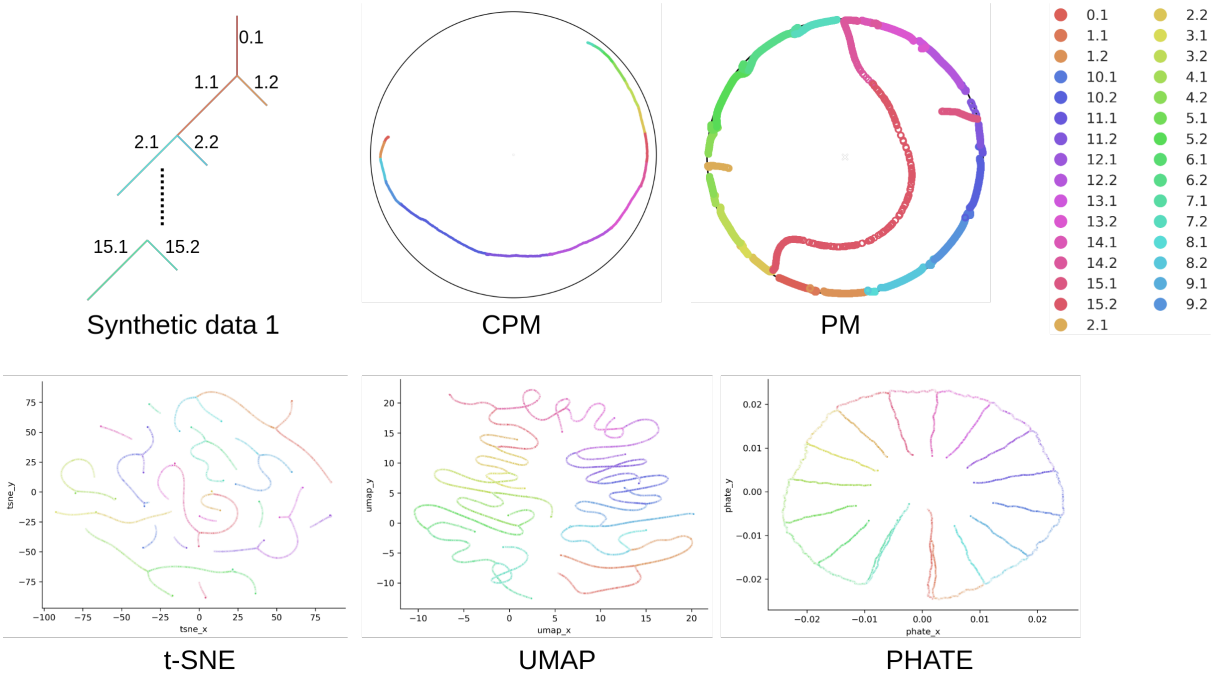

Figure 12: Visualization of embeddings for synthetic data 1, a deep binary tree with 15 hierarchical levels, where each node branches into two children.

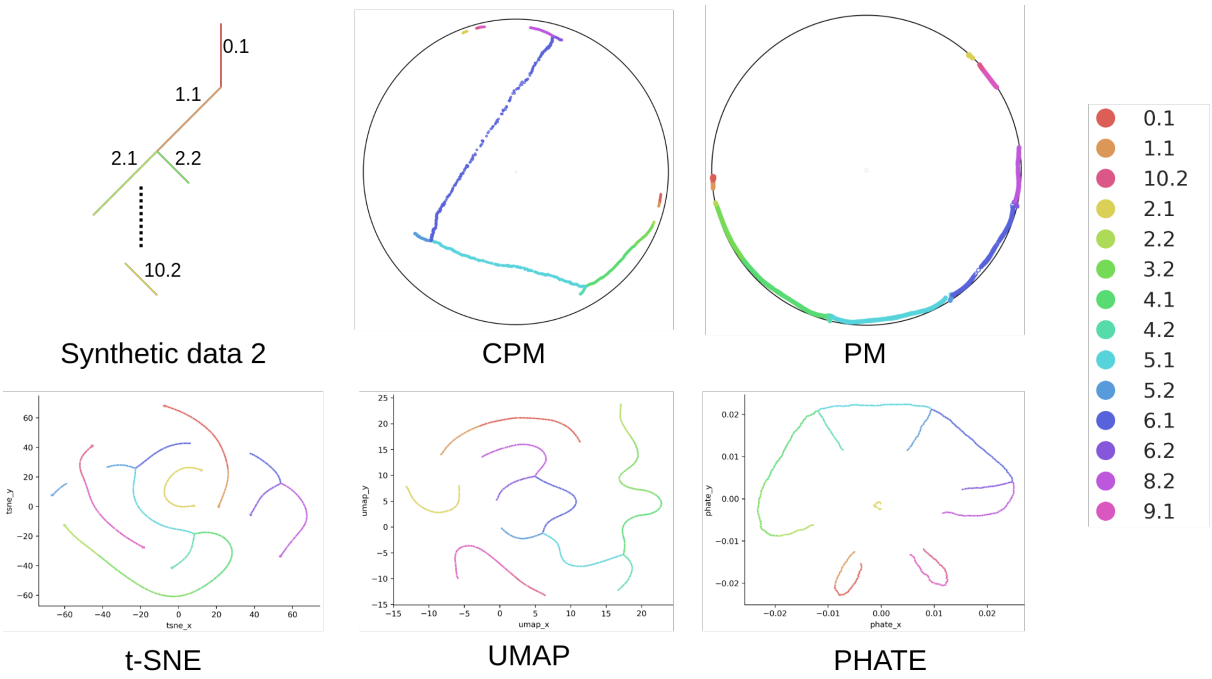

Figure 13: Visualization of embeddings for synthetic data 2, an imbalanced binary tree with 10 levels, in which each node also has two children, but branches are sampled randomly to introduce structural irregularity.

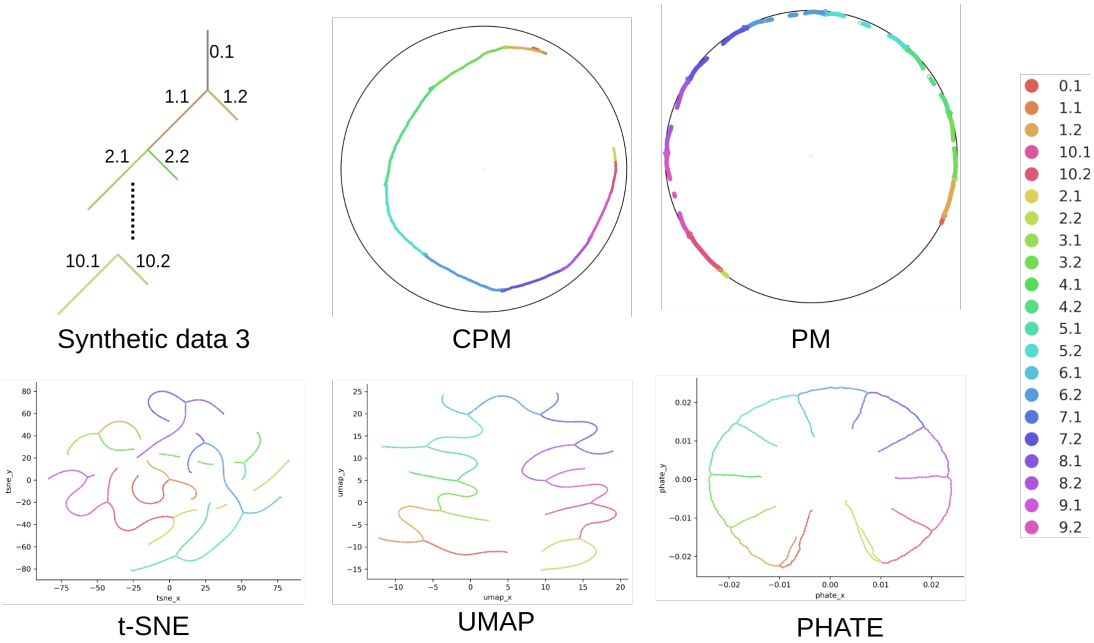

Figure 14: Visualization of embeddings for synthetic data 3, a balanced binary tree with 10 levels, each node branching into two children, with no randomness in structure.

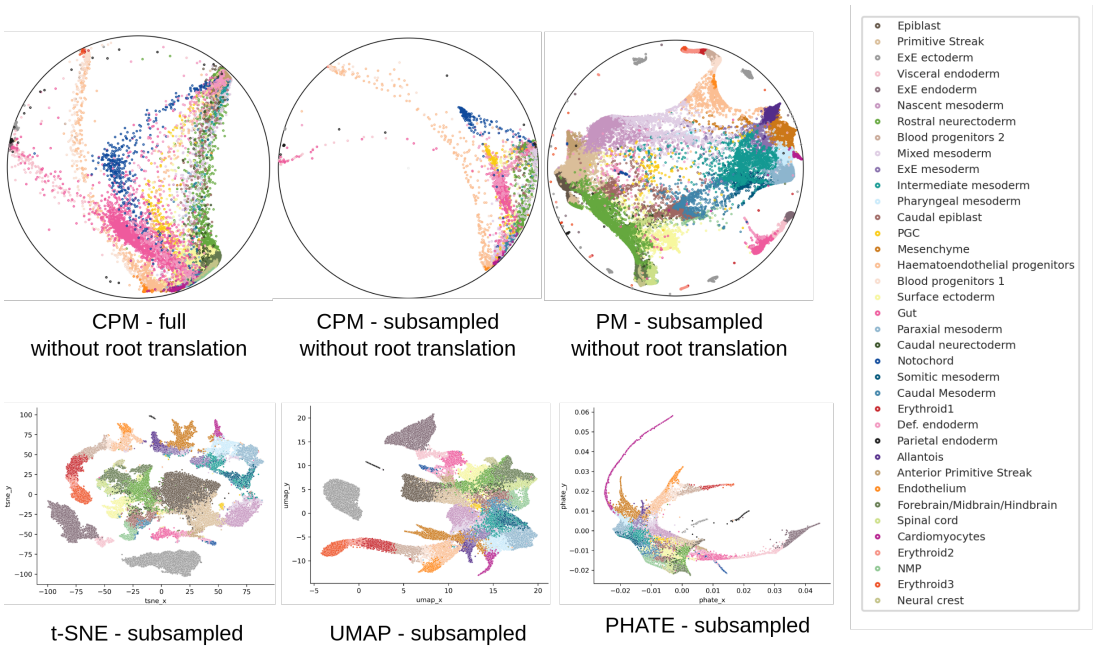

Figure 15: Visualization of embeddings for the subsampled mouse gastrulation dataset, comprising 58,156 cells across nine developmental stages. For Contrastive Poincaré Maps, the visualization for the full dataset, comprising 116,312 cells is also provided.

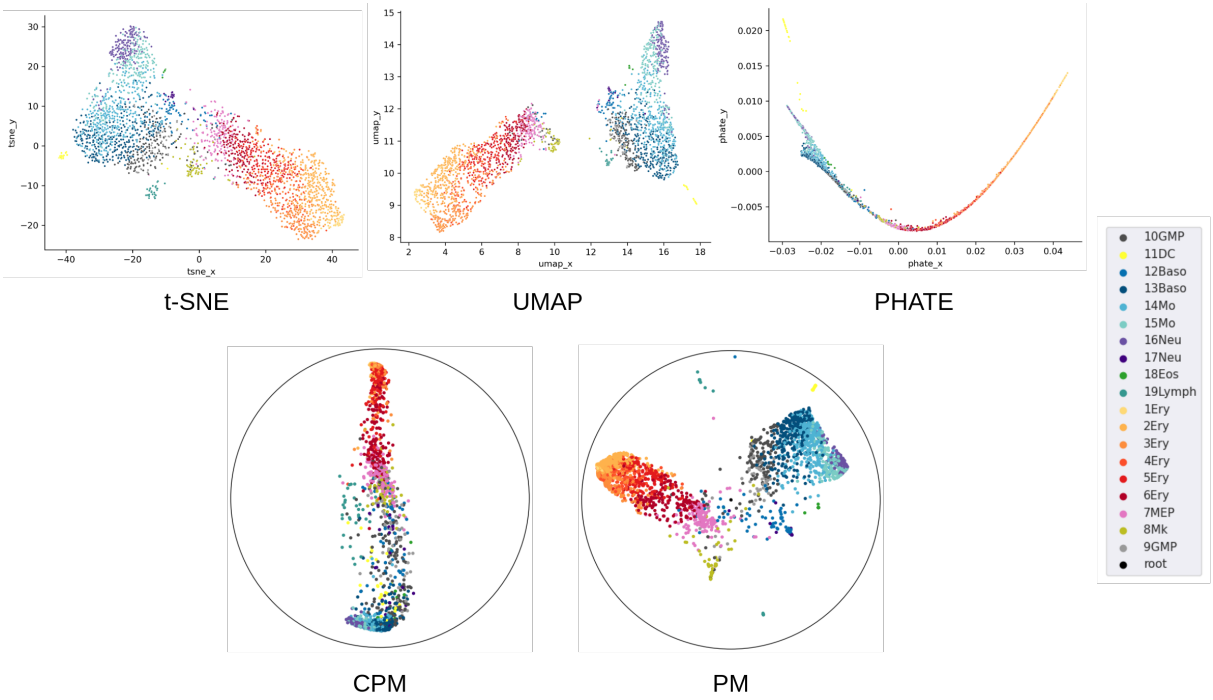

Figure 16: Visualization of embeddings for the Mice haematopoiesis dataset, comprising 2,730 myeloid progenitor cells.

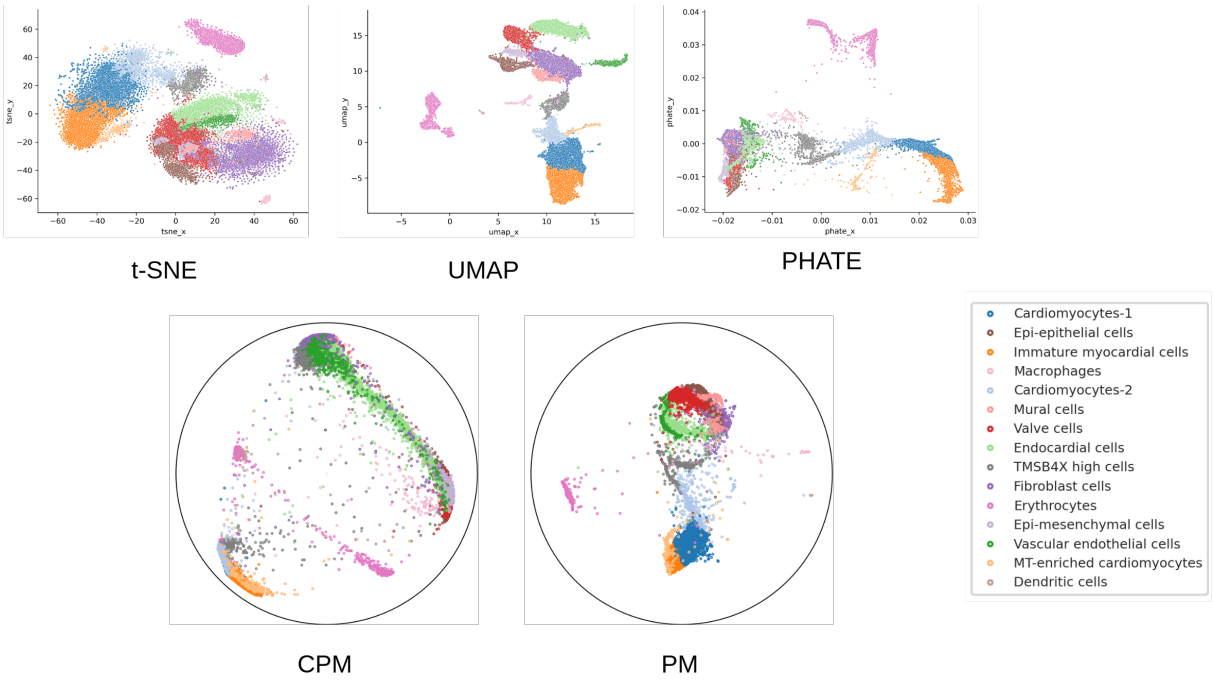

Figure 17: Visualization of embeddings for the Chicken cardiogenesis dataset, comprising 22,315 cells from multiple lineages across four developmental stages.

E Additional results: Mouse gastrulation

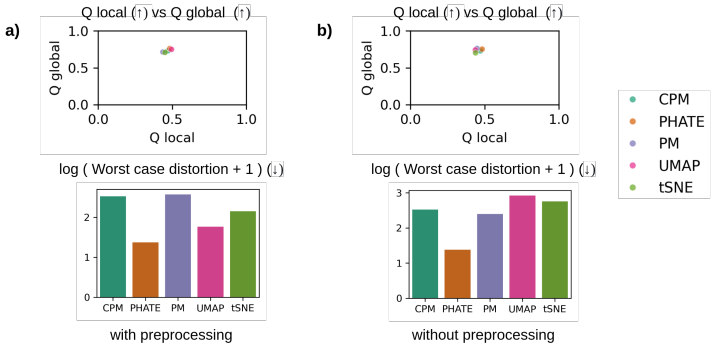

Figure 18: **Quantitative evaluation on subsampled mouse gastrulation data.** Performance of all embedding methods evaluated on a subsample of 58,156 cells from the full mouse gastrulation dataset. Due to computational constraints associated with pairwise distance calculations on the full dataset, metrics were computed on the subsample. Additional results with PCA-based dimensionality reduction are included for all methods except Contrastive Poincaré Maps.

F Additional results: Translation to single nucleus RNA sequencing dataset

We analyze single-nucleus RNA sequencing (snRNA-seq) data from mouse brain generated by Russell et al. (2024), who developed the Slide-tags method to achieve spatially resolved transcriptomics at single-nucleus resolution. This dataset comprises of spatially resolved transcriptomes from 4,584 nuclei, with a median of 4,594 UMIs per nucleus. These data were derived from a $7mm^2$ sagittal section of the mouse brain at embryonic day 14 (E14). While previous results were primarily focused on scRNA-seq datasets, we use this dataset to demonstrate capabilities of CPM in resolving cellular hierarchies from snRNA-seq data as well. Specifically, we use it to show that CPM can resolve localized differentiation hierarchies within a single embryonic stage.

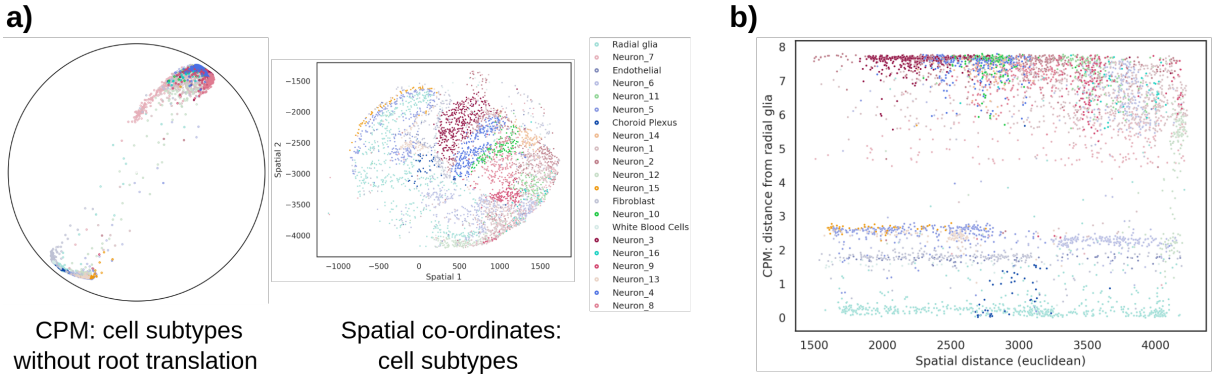

Figure 19: **Mouse brain development at E14** (a) Embedding learned by Contrastive Poincaré Maps and the corresponding spatial coordinates from the dataset, colour-coded by cell subtype. (b) Distances from radial glia in the embedding plotted against spatial distances. The plot reveals discrete tiers of differentiation, indicating that Contrastive Poincaré Maps captures global transcriptional hierarchies, independent of spatial proximity.

Figure 19a presents the embedding generated by CPM, alongside the original spatial coordinates from the dataset. Three key observations emerge from these results: (i) CPM partitions the data into two principal clusters - one comprising radial glia, choroid plexus, white blood cells, endothelial cells, fibroblasts, and a subset of neurons, and the other consisting predominantly of neurons; (ii) Inferred developmental trajectories place Neuron 13, Neuron 5, and Neuron 15 closer to progenitor populations (Radial glia). (iii) Neuron 6 and Neuron 12 appear to originate early in the differentiation process but remain transcriptionally abundant at both early and late stages (see Fig. 20), indicating sustained expression across developmental time as inferred from CPM embedding. Figures 21 and 22 illustrate the quantitative and qualitative evaluation of all embedding methods.

Figure 19b shows a scatter plot of embedding distances from radial glia versus spatial distances provided in the dataset. The lack of correlation between these two distance measures highlights that CPM captures global differentiation hierarchies based on transcriptional cell states, rather than spatial proximity. Specifically, the plot reveals discrete waves or tiers of cell differentiation, as inferred from the embedding. While spatial distances reflect the physical adjacency of cell types within the tissue, the embedding distances generated by CPM reflect their relative maturation along the differentiation trajectory.

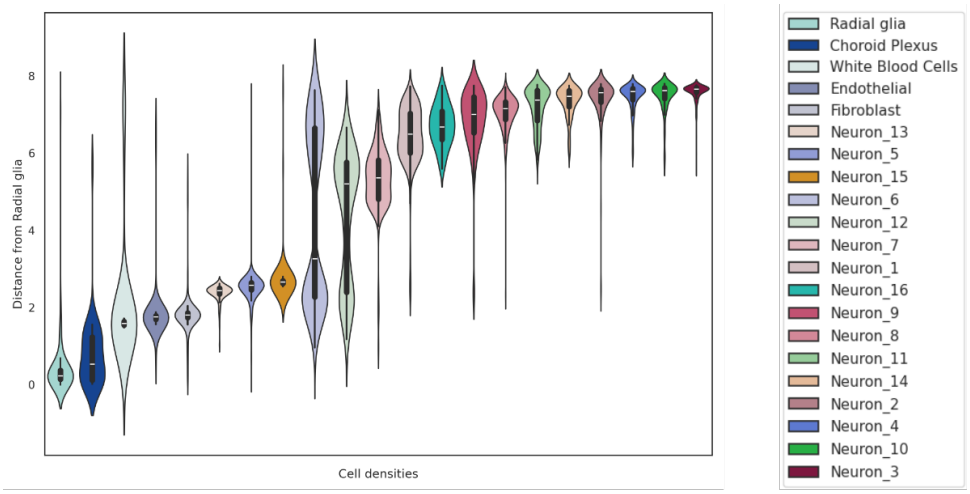

Figure 20: **Cell densities in mouse brain development** Cell subtype densities in the mouse brain development dataset visualized as a function of distance from radial glia in the Contrastive Poincaré Maps embedding. This analysis supports inference of cell ordering along developmental trajectories while cross-validating with local cell density distributions.

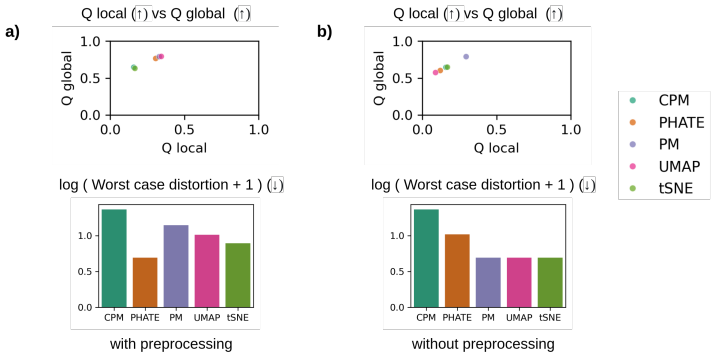

Figure 21: **Quantitative evaluation on mouse brain development dataset.** Performance of all embedding methods on mouse brain development dataset evaluated (a) with and (b) without PCA-based dimensionality reduction. The results of Contrastive Poincaré Maps (CPM) in both (a) and (b) are without PCA-based dimensionality reduction.

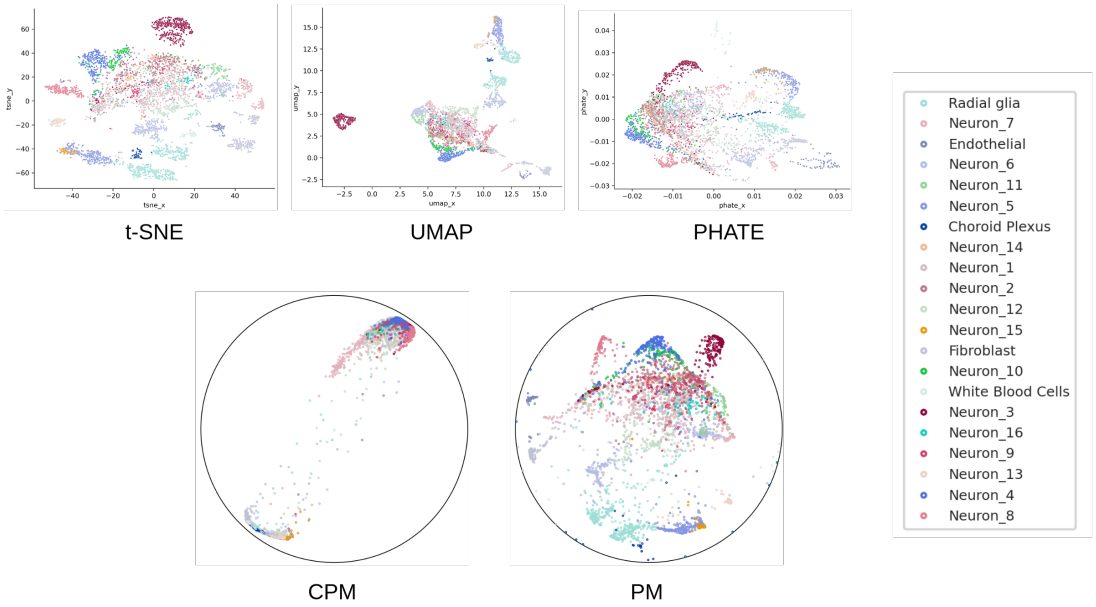

Figure 22: Visualization of embeddings for the mouse brain development dataset, comprising single-nucleus RNA sequencing data from 4,853 nuclei at embryonic day E14.

G Additional results: Analysis of gene expression profiles

This section presents gene expression profiles of key marker genes across all real datasets. Expression values are min–max normalized to facilitate comparison and overlaid on the CPM embedding. The overlays reveal transcriptional gradients and global hierarchical organization of cell states, reflecting underlying differentiation trajectories.

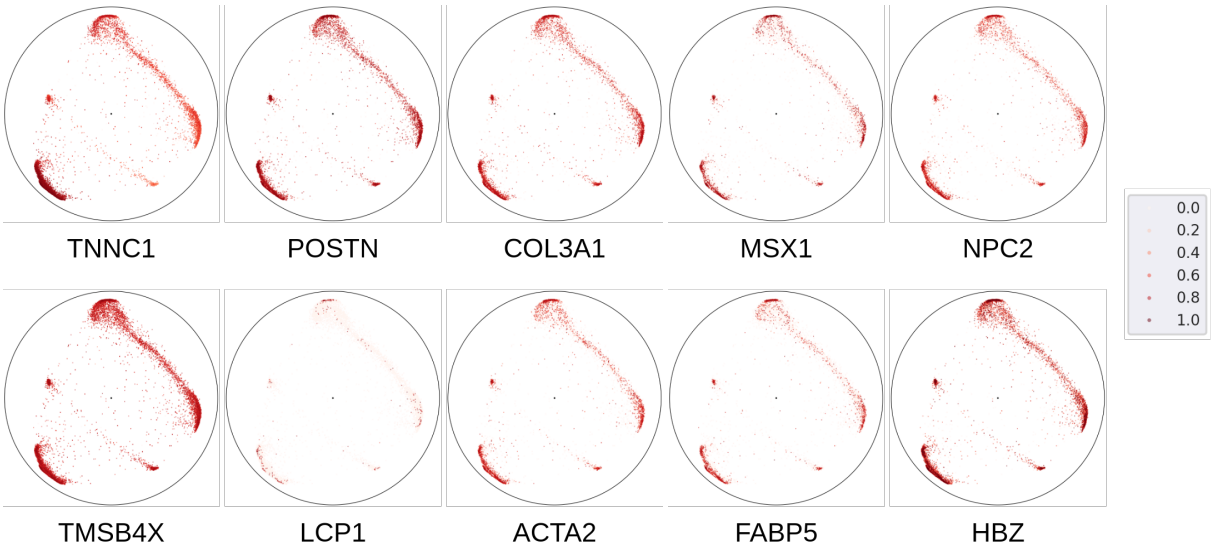

Figure 23: **Gene expression patterns overlaid on CPM embeddings** Min–max normalized expression profiles of representative marker genes for the chicken cardiogenesis dataset, visualized on CPM embeddings.

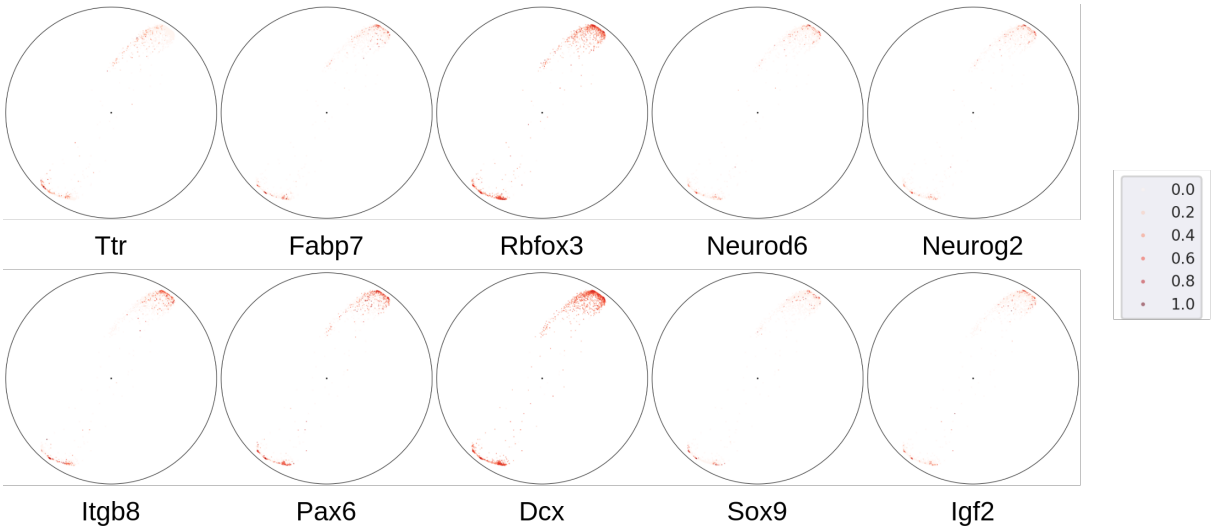

Figure 24: **Gene expression patterns overlaid on CPM embeddings** Min–max normalized expression profiles of representative marker genes for the mouse brain development dataset, visualized on CPM embeddings.

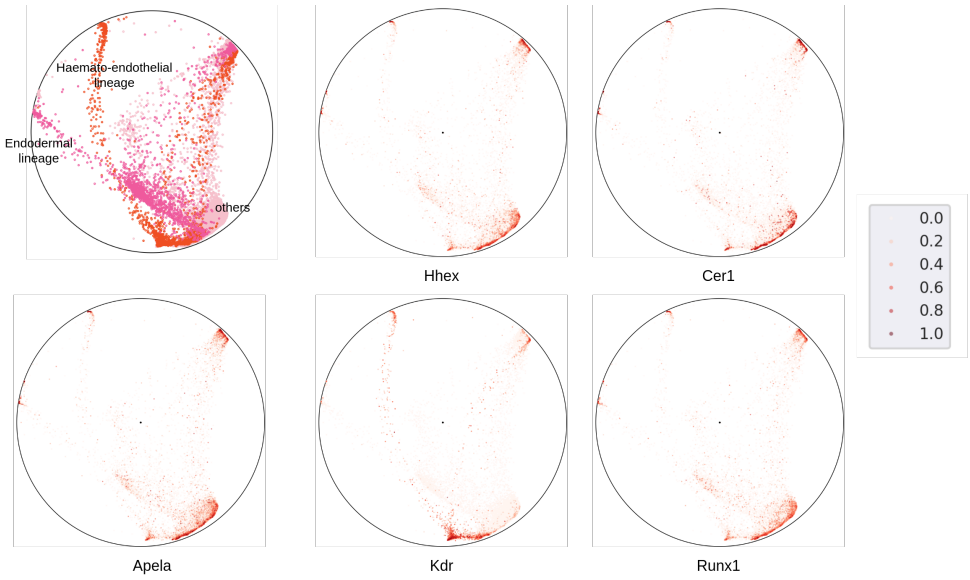

Figure 25: **Gene expression patterns overlaid on CPM embeddings** Min–max normalized expression profiles of representative marker genes for the mouse gastrulation dataset, visualized on CPM embeddings.

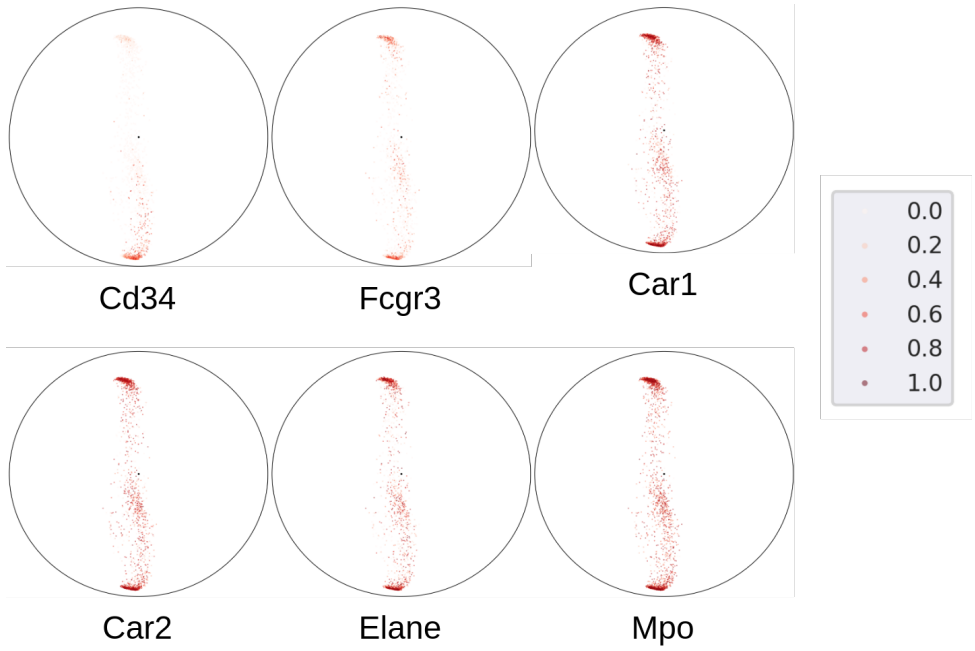

Figure 26: **Gene expression patterns overlaid on CPM embeddings** Min–max normalized expression profiles of representative marker genes for the mouse haematopoeisis dataset, visualized on CPM embeddings.