PHATE: A Dimensionality Reduction Method for Visualizing Trajectory Structures in High-Dimensional Biological Data

Kevin R. Moon,[1,2†] David van Dijk,[5†] Zheng Wang,[4†] William Chen,[1]
Matthew J. Hirn,[6,7] Ronald R. Coifman,[2] Natalia B. Ivanova,[4‡**]
Guy Wolf,[2‡] Smita Krishnaswamy[1,3‡*]

[1]Departments of Genetics; [2]Applied Mathematics Program;
[3]Department of Computer Science;
[4]Yale Stem Cell Center, Department of Genetics,
Yale University, New Haven,CT,USA
[5]Computational Biology Program, Memorial Sloan-Kettering
Cancer Center, New York, NY, USA
[6]Department of Computational Mathematics, Science and Engineering;
[7] Department of Mathematics, Michigan State University,
East Lansing, MI, USA

[*]Corresponding author. E-mail: smita.krishnaswamy@yale.edu
Address: 333 Cedar St, New Haven, CT 06510, USA
[**]Correspondence for experiments. E-mail: natalia.ivanova@yale.edu

[†] These authors contributed equally. [‡] These authors contributed equally.

Abstract

In recent years, dimensionality reduction methods have become critical for visualization, exploration, and interpretation of high-throughput, high-dimensional biological data, as they enable the extraction of major trends in the data while discarding noise. However, biological data contains a type of predominant structure that is not preserved in commonly used methods such as PCA and tSNE, namely, branching progression structure. This structure, which is often non-linear, arises from underlying biological processes such as differentiation, graded responses to stimuli, and population drift, which generate cellular (or population) diversity. We propose a novel, affinity-preserving embedding called PHATE (Potential of Heat-diffusion for Affinity-based Trajectory Embedding), designed explicitly to preserve progression structure in data.

PHATE provides a denoised, two or three-dimensional visualization of the complete branching trajectory structure in high-dimensional data. It uses heat-diffusion processes, which naturally denoise the data, to compute cell-cell affinities. Then, PHATE creates a diffusion-potential geometry by free-energy potentials of these processes. This geometry captures high-dimensional trajectory structures, while enabling a natural embedding of the intrinsic data geometry. This embedding accurately visualizes trajectories and data distances, without requiring strict assumptions typically used by path-finding and tree-fitting algorithms, which have recently been used for pseudotime orderings or tree-renderings of cellular data. Furthermore, PHATE supports a wide range of data exploration tasks by providing interpretable overlays on top of the visualization. We show that such overlays can emphasize and reveal trajectory end-points, branch points and associated split-decisions, progression-forming variables (e.g., specific genes), and paths between developmental events in cellular state-space. We demonstrate PHATE on single-cell RNA sequencing and mass cytometry data pertaining to embryoid body differentiation, IPSC reprogramming, and hematopoiesis in the bone marrow. We also demonstrate PHATE on non-single cell data including single-nucleotide polymorphism (SNP) measurements of European populations, and 16s sequencing of gut microbiota.

1 Introduction

Biological data are often developmental in nature and can be characterized by various types of progressions. In particular, progression is inherent to single-cell data since all human body cells arise from a single oocyte, which differentiates into the various tissues and subtypes. For example, progression is present in directed differentiation of embryonic stem cells, which has recently shown promise for regenerative medicine. Additionally, cells in many areas of the body are actively differentiating or progressing in response to signals. For instance, bone marrow cells are constantly differentiating from hematopoetic stem cells into myeloid and lymphoid cells. Cells in the embryo can undergo a progression known as the epithelial-to-mesenchymal transition, which turns epithelial cell types into free-floating mesenchymal cell types (a process hijacked by cancer).

Progression is also inherent to other biological datatypes. For example, gut bacterial species in patients with autoimmune conditions can show progression based on the extent of the underlying disease. Population genetic data can show progression in genotypes based on population drift and admixture events.

There has recently been an explosion in high-throughput technologies that can measure such progressions in biology. Examples include single-cell RNA-sequencing (scRNAseq), mass cytometry, SNP arrays, and microbiome sequencing. New snapshot single-cell technologies (such as those in mass cytometry or scRNAseq) can capture cells in all phases of these progressions.

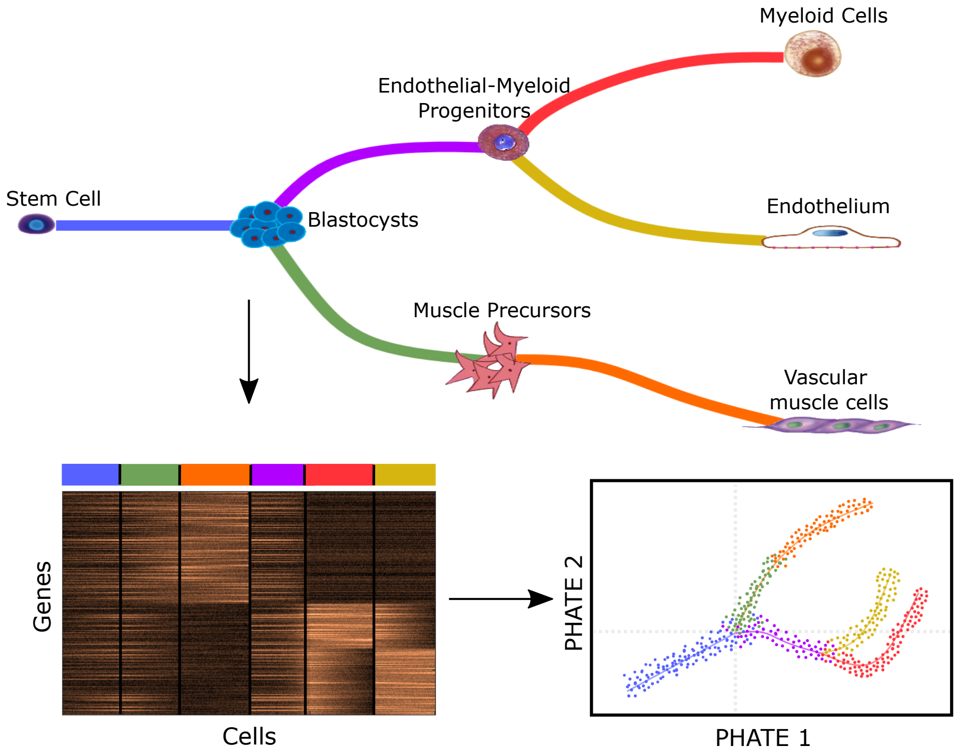

Figure 1: Conceptual figure demonstrating the progression of stem cells into different cell types and the corresponding high dimensional single-cell measurements (e.g., mass cytometry or scRNAseq). PHATE embeds the progression structure within the high dimensional data into lower dimensions (e.g., 2D or 3D) for visualization. The trajectories and branches can then be analyzed to extract biological meaning.

Other technologies, such as SNP arrays and microbiome sequencing, can measure progression between patients.

Many of these high-throughput technologies provide high dimensional data (e.g., gene expression levels for thousands of genes in scRNAseq data), which can be used to characterize the biological progressions in great detail. For instance visualizing different cell-fates in the data in terms of genes that increase or decrease expression along trajectories is key to understanding what drives certain paths. However, the high dimensional and noisy nature of the data also makes it difficult to extract or visualize the progression (see Fig. 1) or to use it for data exploration.

Data dimensionality reduction methods such as PCA, and more recently, tSNE [1] have been used for biological data visualization. However, these methods do not address the urgent need in biology to visualize and understand high-dimensional progression or branching trajectory structures that often occur as a dominant underlying pattern in biological systems.

To address this need, we present a new dimensionality reduction technique to optimally characterize, organize, and visualize biological data given its highly non-linear structure, noise (both biological and technical), and continuous progressive nature. We call our new method ***PHATE*** (***Potential of Heat-diffusion for Affinity-based Trajectory Embedding***). PHATE constructs a non-linear embedding of high dimensional data that simultaneously denoises the data and emphasizes the continuous nature of any underlying progressions and trajectories. PHATE naturally uncovers branching progression structures in the data in very low (i.e., two or three) dimensions to enable visualization. We show that this method outperforms existing methods in terms of revealing correct underlying structure in a low dimensional visualization. Additionally, PHATE has advantages over tree-rendering techniques that initially cluster the data and then artificially construct the data as a tree (methods such as Monocle2 [2] or SPADE [3]): PHATE is a true dimensionality reduction method that preserves heat-diffusion potential distances such that trajectory structure is naturally and accurately emphasized. Therefore, PHATE is stable and robust and will not provide a different rendering at each run. Additionally, PHATE embeddings can be colored by local intrinsic dimensionality to reveal branch points, eigencentrality to reveal endpoints, and various genes to reveal the progression of gene expressions along branches. We demonstrate the utility of PHATE on a wide variety of biological datasets that contain large sample sizes, primarily scRNAseq and mass cytometry (CYTOF) data. We also show results on gut microbiome data, and on SNP (single-nucleotide polymorphism) population genetics data to emphasize the generality of our visualization on any high dimensional data matrix. In addition, we describe methods for extracting quantitative information from PHATE such as branch point and branch identification. This can then be used to identify genes that correlate with branches to derive biological meaning from PHATE. We note that PHATE complements methods that extract pseudo-time orderings from data, including Wanderlust [4], and Wishbone [5] as they can be run on top of PHATE dimensions.

2 The PHATE Algorithm

The development of PHATE was inspired by Word2vec [6], Glove [7] and other algorithms that find low-dimensional metric embeddings of words. These methods take advantage of the observation that meaningful representations of words should not consider them individually, but rather as parts of a phrase or a sentence whose progression develops semantic notions. They use the structure provided by input text to define and associate a context with each word, and in turn, identify similarities between words by their contexts. Then, they construct an embedding of words into a vector space by ensuring that the proximities between embedded vectors correlate with similar textual contexts. Surprisingly, the relations uncovered by such context-based metric embedding is not only proximal, but it has been shown that directionality in the embedded space uncovers semantic progression between words. For example, specific directions identify gender relations (e.g., male-female and king-queen), geographical relations (e.g., Spain-Madrid, Italy-Rome, and Germany-Berlin), or even grammatical conjugations (e.g., walked-walking or swam-swimming).

Biological data often consists of developmental progressions that can be continuously observed in cells. For example, cells gradually change state during the course of differentiation processes. Therefore, a cascade of local cell-cell similarities (e.g., nearest neighbor affinities) can be used to define a developmental context that reveals differentiation pathways, and thus expresses cells as parts of such progression. Unlike text processing, in our case the input data is given in unstructured form, and therefore the proposed PHATE method must infer the context of cells in order to utilize it for embedding and visualization purposes. First PHATE computes local affinities between cells, and then these affinities are used to define transitional probabilities and propagate them via a Markovian diffusion process over the data. This causes the data to separate and contract onto diffusion trajectories, which are spread among numerous orthogonal directions identified by the eigenstates of the diffusion process.

To stabilize the diffusion trajectories and allow their embedding in a low-dimensional (most importantly - easily visualizable) space, we transform the diffusion transitional probabilities into a novel, localized heat potential representation. The context of each cell in the data is then represented by the potential of the heat it propagates to other cells. These heat-potential contexts are embedded into a two dimensional space using non-metric multi-dimensional scaling (MDS), which preserves monotone relations between potential distances. In other words, the data is organized by preserving monotone ordering of developmental context variations; thus, it visually emphasizes progression branches and trajectories.

We demonstrate PHATE on a synthetically generated dataset that uses diffusion limited aggregation [11] to generate an artificial tree-like structure. This data was generated to have 20 branches in 100 dimensions and 100 data points per branch. We added noise to the tree (see Methods) and then compared the PHATE embedding to PCA, tSNE, and diffusion maps (DM) in Fig. 2. The PCA embedding preserves some of the global structure of the data. However, the local information is lost due to the noise and the nonlinear structure of the data and thus the structure appears fuzzy. The tSNE embedding preserves some of the local structure branching structure but loses all global structure as it shatters the trajectories into clusters. The DM embedding preserves some global progression structure. However, it tends to put each progression into a different dimension and does not result in a low-dimensional embedding. In contrast, the PHATE embedding is best at finding both the global and local progression structures and preserving them in low dimensions.

We perform a similar comparison on several single-cell biological datasets in Fig. 3. The datasets used include: 1. Developing mouse bone marrow cells, enriched for the myeloid and erythroid lineages, which were measured with the MARS-seq single cell RNA-sequencing technology [8]; 2. Developing mouse bone marrow cells, enriched for lymphoid lineages, as measured via mass cytometry [9]; 3. Mass cytometry data showing iPSC reprogramming of mouse embryonic fibroblasts [10]. PHATE is the only method designed to emphasize and preserve trajectory structure in the data. The biological datasets represent differentiating processes within the body, and hence visualizing progression is key to understanding the structure of this data.

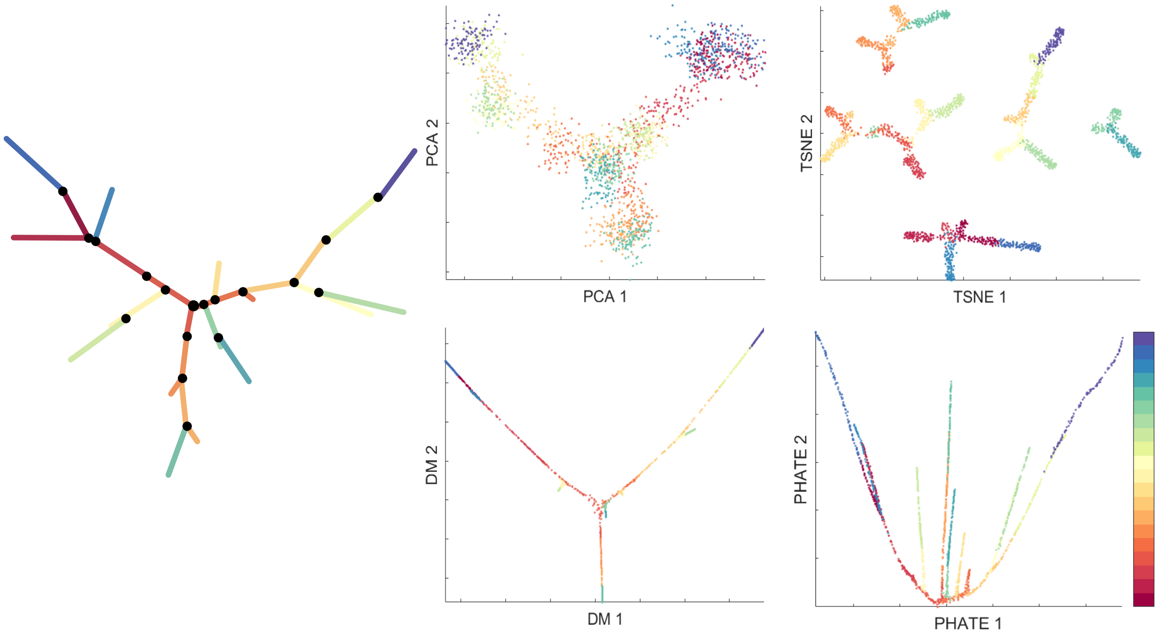

Figure 2: PHATE applied to artificial tree branching data with 20 branches in 100 dimensions and 100 data points per branch. (Left) A 2D drawing of the noiseless artificial tree colored by branch. (Right) A comparison of the PHATE embedding to PCA, tSNE, and diffusion maps (DM) with data points colored by branch. The scale for the DM and PHATE embeddings is $t = 30$. The PHATE embedding is best at finding the global structure of the data while simultaneously distinguishing more of the smaller branches from the global structure.

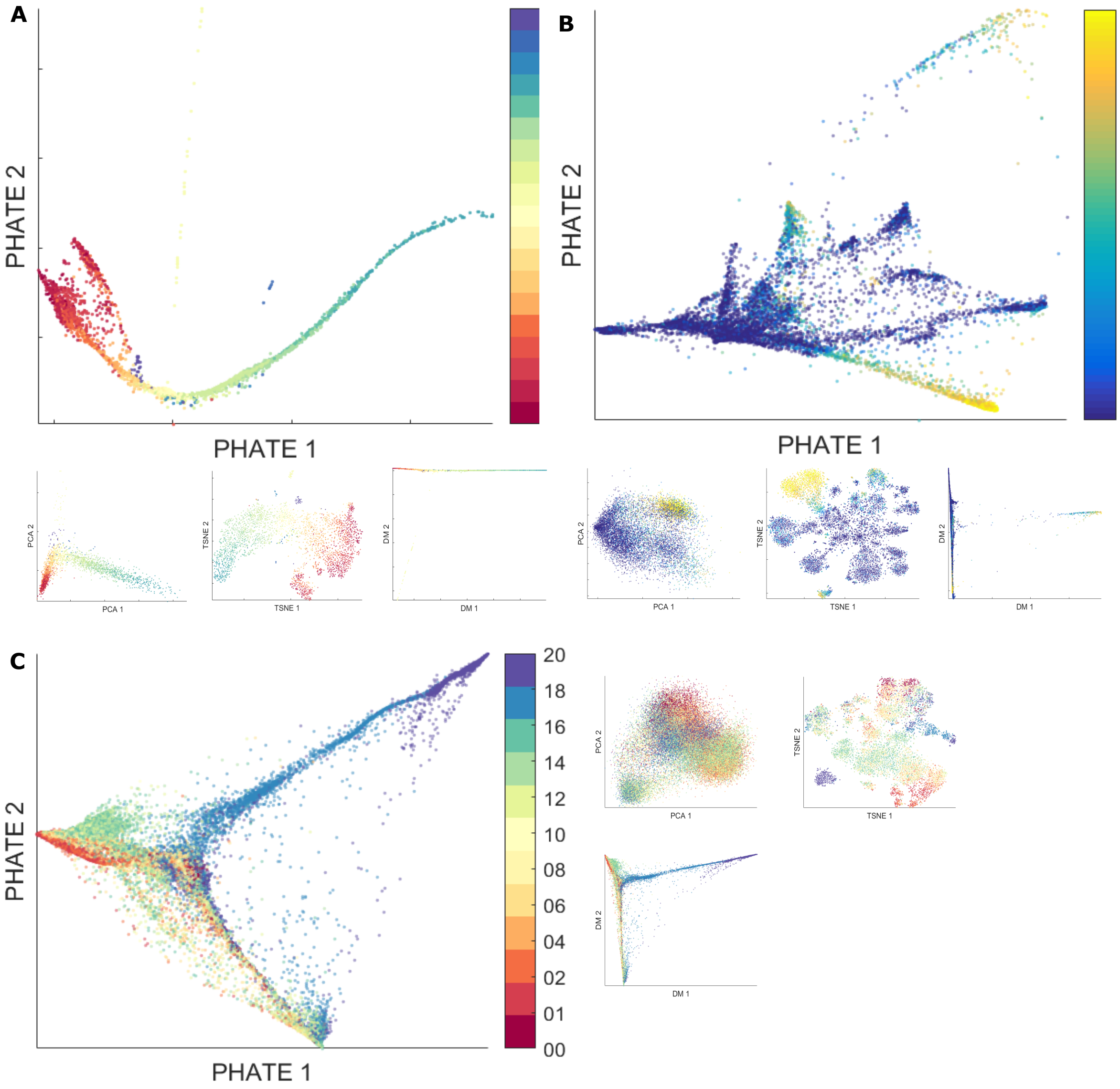

Figure 3: Comparison of PCA, tSNE, diffusion maps (DM), and the PHATE embeddings for various data sets. PHATE is the only method designed to emphasize and preserve trajectory structure in data for visualization. (**A**) Mouse bone marrow scRNAseq data colored by cell type as identified in [8]. See Fig. 10 for a legend. The scale for DM and PHATE is $t = 40$. (**B**) Bone marrow mass cytometry data [9] subsampled at $N = 10000$ points and colored by CD4 expression level. The scale for DM and PHATE is $t = 100$. (**C**) iPSC CyTOF data [10] subsampled at $N = 50000$ points and colored by sample time. The scale for DM and PHATE is $t = 250$.

2.1 Manifold and Diffusion Geometry Data Models

In order to establish an abstract geometric model for the types of data that are suitable for PHATE, we consider two properties: 1. Development occurs incrementally, as an aggregation of many small modifications, and 2. There are a limited number of possible outcomes from each incremental modification. These properties, which are valid in cellular developmental progression, indicate that instantaneous progression can be captured and expressed by locally low dimensional neighborhoods of observed cells. Progression tracks can thus be modeled geometrically by smoothly varying data patches defined by such neighborhoods. This collection essentially constitutes a mathematical manifold model for the geometry of a progression track. Furthermore, such manifolds have a low intrinsic dimension, even if curvature and noise forces them to span a high dimensional volume in the collected feature space. Finally, in the case of cellular progression, progression tracks form trajectories, with a small number of "branching points", where progression splits into several directions. Therefore, in this case it is useful to model the data as a collection of intrinsically one-dimensional manifolds (i.e., curves) that cross each other in branching points.

It has been shown in several works (e.g., [12, 13]) that manifold geometries are closely related to heat diffusion, modeled by the differential heat equation, on the one hand, and to differential Laplace-Beltrami operators on the other hand. Indeed, solutions of the heat equation over a manifold capture its intrinsic properties, while providing embeddings, affinities, and distance metrics that capture intrinsic manifold relations. It has further been shown that these can be robustly discretized for empirical observations that correlate with hidden (or latent) manifold models, e.g., by considering diffusion maps embedding of the data [14–16]. The embedding obtained by PHATE extends these results by considering an underlying geometry consisting of multiple one-dimensional manifolds (i.e., trajectory curves) that cross each other, while alleviating boundary-condition instabilities to maintain low dimensionality of the embedded space. We note that the trajectory structure is not artificially generated in our case, but rather it is expected to be dominant (albeit latent or hidden) in the data. Therefore, the PHATE visualization will only show trajectory structures when data fits such a geometry; otherwise, other (e.g., cluster) patterns will be expressed in the PHATE visualization.

2.2 Overview of the PHATE Algorithm

The main steps for obtaining the proposed embedding are described in Alg. 1. PHATE involves computing a localized Markov transition matrix (henceforth called a *diffusion operator*) between cells (or samples). This operator is computed by first computing local affinities between points and then normalizing the affinities such that they become transition probabilities between cells. Then we power or diffuse the matrix to obtain longer-range, cleaner connections between cells. Then we transform these transition probabilities into the heat-potential context. Finally, we embed the resultant matrix with non-metric MDS for visualization in low dimensions. These steps are demonstrated in Fig. 4 by a block diagram, which shows the main

Algorithm 1: The PHATE algorithm

```
Input: Data matrix $X$, neighborhood size $k$, locality scale $\alpha$
Output: The PHATE embedding $Y$
1: $D \leftarrow$ compute pairwise distance matrix from $X$
2: Compute the $k$-nearest neighbor distance $\varepsilon_k(x)$ for each column $x$ of $X$
3: $K_{k,\alpha} \leftarrow$ compute local affinity matrix from $D$ and $\varepsilon_k$ (see Eq. 3)
4: $P \leftarrow$ normalize $K_{k,\alpha}$ to form a Markov transition matrix (diffusion operator; see Eq. 2)
5: $t \leftarrow$ compute time scale via Von Neumann Entropy (see Eq. 7)
6: Diffuse $P$ for $t$ time steps to obtain $P^t$
7: Compute potential representations: $U_t \leftarrow -\log(P^t)$
8: $D_{U,t} \leftarrow$ compute potential distance matrix from $U_t$ (see Def 1)
9: $Y \leftarrow$ apply nonmetric MDS of $D_{U,t}$ to embed in $\mathbb{R}^2$ (or $\mathbb{R}^3$)
```

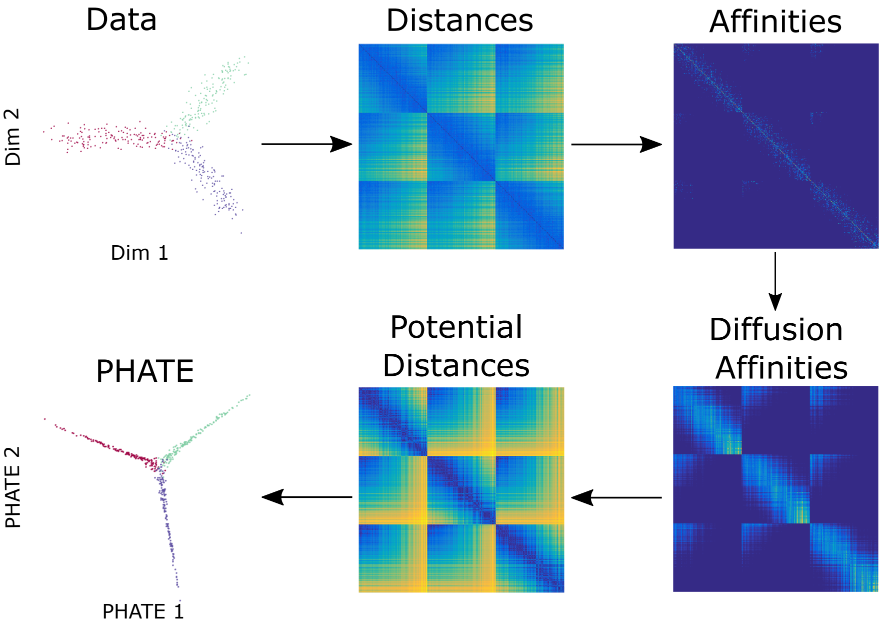

Figure 4: Block diagram demonstrating the main matrices computed by the PHATE algorithm (Alg. 1) when embedding noisy tree structure with 3 branches from $\mathbb{R}^{12}$ into $\mathbb{R}^2$.

matrices computed by PHATE to embed an artificially generated tree structure. Once the two- or three-dimensional embedding is constructed, it can be visualized to allow intelligible exploration and determination of branching and trajectory structures. We note that PHATE is different from a diffusion map in that it does not eigendecompose the powered diffusion operator directly but rather uses a distance preserving embedding of cells re-represented by their potential heat distances to other cells. This has the affect of collecting trajectories in low dimensions rather than spreading them out into individual dimensions like diffusion maps.

The following sections provide detailed explanations regarding each of the steps in the algorithm. Furthermore, we also propose and describe methods for automatically annotating the provided visualization by extracting branching and trajectory information from the embedding.

2.3 The Diffusion Operator

PHATE is based on constructing a diffusion geometry to learn and represent the shape of the data [14–16]. This construction is based on computing local similarities between data points, and then *walking* or *diffusing* through the data using a Markovian random-walk diffusion process to infer more global relations. The local similarities between points are computed by first computing Euclidean distances and then transforming the distances into similarities, typically via a Gaussian kernel. This kernel has the advantage of emphasizing local distances and decaying relatively rapidly after one standard deviation.

Let $\mathcal{X} = \{x_1, \dots, x_N\} \subset \mathbb{R}^d$ be a dataset sampled i.i.d. from a probability distribution $p : \mathbb{R}^d \to [0, 1]$ (with $\int p(x)dx = 1$) that is essentially supported on a low dimensional manifold $\mathcal{M}^m \subseteq \mathbb{R}^d$ with $m \ll d$. The classic diffusion geometry proposed in [14] is based on first defining a notion of local neighborhoods in the data. A popular locality notion is given by a Gaussian kernel $k_\varepsilon(x, y) = \exp(-\|x - y\|^2/\varepsilon)$ that quantifies similarities between points based on Euclidean distances. The bandwidth ε determines the radius (or spread) of neighborhoods captured by this kernel. The kernel is then normalized with the row-sums

$$\nu_\varepsilon(x) = \|k_\varepsilon(x, \cdot)\|_1 = \sum_{j=1}^{N} k_\varepsilon(x, x_j) \tag{1}$$

resulting in a $N \times N$ row-stochastic matrix

$$[P_\varepsilon]_{(x,y)} = \frac{k_\varepsilon(x, y)}{\nu_\varepsilon(x)}, \qquad x, y \in \mathcal{X}\,. \tag{2}$$

The matrix P_ε is a Markov transition matrix where the probability of moving from x to y in a single time step is given by $\Pr[x \to y] = [P_\varepsilon]_{(x,y)}$.

2.3.1 The alpha-decaying kernel and adaptive bandwidth

When applying the diffusion map framework to data, the choice of the kernel K and bandwidth ε plays a key role in the results. In particular, choosing the bandwidth corresponds to a tradeoff between encoding global and local information in the probability matrix P_ε. If the bandwidth is small, then single-step transitions in the random walk using P_ε are largely confined to the nearest neighbors of each data point. In biological data, trajectories between major cell types may be relatively sparsely sampled. Thus, if the bandwidth is too small, then the neighbors of points in sparsely sampled regions may be excluded entirely and the trajectory structure in the probability matrix P_ε will not be encoded. Conversely, if the bandwidth is too large, then the resulting probability matrix P_ε loses local information as $[P_\varepsilon]_{(x,\cdot)}$ becomes more uniform for all $x \in \mathcal{X}$, which may result in an inability to resolve different trajectories. Here, we use an adaptive bandwidth that changes with each point to be equal to its kth nearest neighbor, along with an α-decaying kernel that controls the rate of decay of the kernel.

The original heuristic proposed in [14] suggests setting ε to be the smallest distance that still keeps the diffusion process connected. In other words, it is chosen to be the maximal 1-nearest neighbor distance in the dataset. While this approach is useful in some cases, it is greatly affected by outliers and sparse data regions. Furthermore, it relies on a single manifold with constant dimension as the underlying data geometry, which may not be the case when the data is sampled from specific trajectories rather than uniformly from a manifold. Indeed, the intrinsic dimensionality in such cases differs between mid-branch points that mostly capture one-dimensional trajectory geometry, and branching points that capture multiple trajectories crossing each other.

This issue can be mitigated by using a locally adaptive bandwidth that varies based on the local density of the data. A common method for choosing a locally adaptive bandwidth is to use the k-nearest neighbor (NN) distance of each point as the bandwidth. A point x that is within a densely sampled region will have a small k-NN distance. Thus, local information in these regions is still preserved. In contrast, if x is on a sparsely sampled trajectory, the k-NN distance will be greater and will encode the trajectory structure. We denote the k-NN distance of x as $\varepsilon_k(x)$ and the corresponding diffusion operator as P_k.

A weakness of using locally adaptive bandwidths alongside kernels with exponential tails (e.g., the Gaussian kernel) is that the tails become heavier (i.e., decay more slowly) as the bandwidth increases. Thus for a point x in a sparsely sampled region where the k-NN distance is large, $[P_k]_{(x,\cdot)}$ may be close to a fully-supported uniform distribution due to the heavy tails. This can be mitigated by using the following kernel

$$K_{k,\alpha}(x,y) = \frac{1}{2}\exp\left(-\left(\frac{\|x-y\|_2}{\varepsilon_k(x)}\right)^\alpha\right) + \frac{1}{2}\exp\left(-\left(\frac{\|x-y\|_2}{\varepsilon_k(y)}\right)^\alpha\right), \tag{3}$$

which we call the α-decaying kernel. The exponent α controls the rate of decay of the tails in the kernel $K_{k,\alpha}$. Increasing α increases the decay rate while decreasing α decreases the decay rate. Since $\alpha = 2$ for the Gaussian kernel, choosing $\alpha > 2$ will result in lighter tails in the kernel $K_{k,\alpha}$ compared to the Gaussian kernel. We denote the resulting diffusion operator as $P_{k,\alpha}$. This is similar to common utilizations of Butterworth filters in signal processing applications [17]. See Fig. 5 for a visualization of the effect of different values of α on the kernel function.

Our use of a locally adaptive bandwidth and the kernel $K_{k,\alpha}$ requires the choice of two tuning parameters: k and α. k should be chosen sufficiently small to preserve local information, i.e., to ensure that $[P_{k,\alpha}]_{(x,\cdot)}$ is not a fully-supported uniform distribution. However, k should also be chosen sufficiently large to ensure that the underlying graph represented by $P_{k,\alpha}$ is sufficiently connected, i.e., the probability that we can *walk* from one point to another within

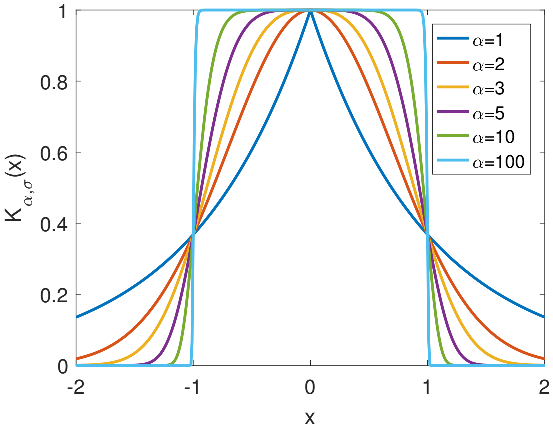

Figure 5: The α-decaying kernel $K_{\alpha,\sigma}(x) = \exp\left(-\left(\frac{|x|}{\sigma}\right)^{\alpha}\right)$ as a function of x for different values of α and $\sigma = 1$. As α increases, $K_{\alpha,\sigma}(x)$ becomes more constant for $x \in (-\sigma, \sigma)$ and the tails of the kernel become lighter (i.e., decay to zero more quickly) for $x \notin (-\sigma, \sigma)$.

the same trajectory in a finite number of steps is nonzero.

The parameter α should also be chosen with k. α should be chosen sufficiently large so that the tails of the kernel $K_{k,\alpha}$ are not too heavy, especially in sparse regions of the data. However, if k is small when α is large, then the underlying graph represented by $P_{k,\alpha}$ may be sparsely connected. Thus we recommend that α be fixed at a large number (e.g. $\alpha \geq 10$) and then k can be chosen to determine the connectivity of the graph. In practice, we find that choosing k to be around 5 and α to be about 10 works well for all the data sets presented in this work.

2.4 Powering the Diffusion Operator

In this section we discuss the motivation for raising the diffusion operator to its t-th power in Alg. 1. To simplify the discussion we use the notation P for the diffusion operator, whether defined with a fixed-bandwidth Gaussian kernel or our adaptive kernel. This matrix is referred to as the diffusion operator, since it defines a Markovian diffusion process that essentially only allows single-step transitions within local data neighborhoods whose sizes depend on the kernel parameters (ε or k and α). In particular, let $x \in \mathcal{X}$ and let δ_x be a Dirac at x, i.e., a row vector of length N with a one at the entry corresponding to x and zeros everywhere else. The t-step distribution of x is the row in P_ε^t corresponding to x:

$$p_x^t \triangleq \delta_x P^t = [P^t]_{(x,\cdot)} . \tag{4}$$

These distributions capture multi-scale (where t serves as the scale) local neighborhoods of data points, where locality is considered via random walks that propagate over the intrinsic manifold geometry of the data.

For appropriate choices of kernel parameters (as describe in previous sections), the diffusion process defined by P is ergodic and it thus has a unique stationary distribution p^∞ that is independent of the initial conditions of the process. Thus $p_x^\infty = p^\infty$ for all $x \in \mathcal{X}$. The stationary distribution p^∞ is the left eigenvector of P with eigenvalue $\lambda_0 = 1$ and can be written explicitly as $\nu/\|\nu\|_1$ with the row-sums from Eq. 1 (possibly adapted to use $K_{k,\alpha}$ from Eq. 3). It can be shown [16] that for fixed-bandwidth Gaussian-kernel diffusion, p^∞ converges asymptotically to the original distribution p of the data as $N \to \infty$ and $\varepsilon \to 0$.

The representation provided by the diffusion distributions p_x^t, $x \in \mathcal{X}$, defines a diffusion geometry with the diffusion distance

$$D^t(x,y) \triangleq \|p_x^t - p_y^t\|_{\ell_2(1/p^\infty)} = \left(\sum_{j=1}^{N} \frac{(p_x^t(x_j) - p_y^t(y_j))^2}{p^\infty(x_j)} \right)^{1/2}, \tag{5}$$

which is given by a weighted ℓ_2 distance between the diffusion distributions originating from the data points x and y. This distance incorporates a comparison between intrinsic manifold regions of the two data points as well as the concentration of data between them, i.e., the difference between the mass distributions.

The diffusion distance at all time scales can be approximated by the Euclidean distance in the diffusion map embedding, which is defined as follows. If the diffusion process is connected, the eigenvalues of P can be indexed as $1 = \lambda_0 > \lambda_1 \geq \cdots \geq \lambda_{N-1} \geq 0$. Let ψ_i and ϕ_i be the corresponding ith left and right eigenvectors of P, respectively. The diffusion map embedding is defined as

$$\Phi^t(x) = (\lambda_1^t \phi_1(x), \lambda_2^t \phi_2(x), \ldots, \lambda_{N-1}^t \phi_{N-1}(x)) \,. \tag{6}$$

The time scale t only impacts the scaling of the embedded coordinates via the powers of the eigenvalues. It can then be shown that $D^t(x,y) = \|\Phi^t(x) - \Phi^t(y)\|_2$.

2.4.1 Choosing the Diffusion Time Scale t with von Neumann Entropy

The diffusion time scale t is an important parameter that affects the embedding. The parameter t determines the number of steps taken in a random walk. A larger t corresponds to more steps compared to a smaller t. Thus, t provides a tradeoff between encoding local and global information in the embedding. The diffusion process can also be viewed as a low-pass filter where local noise is smoothed out based on more global structures. The parameter t determines the level of smoothing. If t is chosen to be too small, then the embedding may be too noisy. On the other hand, if t is chosen to be too large, then some of the signal may be smoothed away.

We choose the timescale t by quantifying the information in the powered diffusion operator with various values of t. We quantify the amount of information in the diffusion operator at time step t by computing the spectral or von Neumann entropy of the powered diffusion operator. The amount of variability explained by each dimension is equal to its eigenvalue in the eigendecomposition of the related (non-Markov) affinity matrix that is conjugate to the Markov diffusion operator. The von Neuman entropy is calculated by computing the Shannon entropy on the normalized eigenvalues of this matrix. Due to noise in the data, this value is artificially high for low values of t, and rapidly decreases as one powers the matrix. Thus, we choose values that are beyond the "knee" of this decrease.

More formally, to choose t, we first note that its impact on the diffusion geometry can be determined by considering the eigenvalues of the diffusion operator, as the corresponding eigenvectors are not impacted by the time scale. To facilitate spectral considerations, we use a symmetric conjugate

$$[A]_{(x,y)} = \sqrt{\nu(x)}[P]_{(x,y)}/\sqrt{\nu(y)}$$

of the diffusion operator P with the row-sums ν. This symmetric matrix is often called the diffusion affinity matrix. We quantify the impact of the time scale t by computing the *Von Neumann Entropy* (VNE) [18,19] of this diffusion affinity. It can be verified that the eigenvalues of A^t are the same as those of P^t, and furthermore these eigenvalues are given by the powers $\{\lambda_i^t\}_{i=1}^{N-1}$ of the spectrum of P. Let $\eta(t)$ be a probability distribution defined by normalizing these (nonnegative) eigenvalues as $[\eta(t)]_i = \lambda_i^t / \sum_{j=0}^{N-1} \lambda_j^t$. Then, the VNE $H(t)$ of A^t is given by the entropy of $\eta(t)$, i.e.,

$$H(t) = -\sum_{i=1}^{N} [\eta(t)]_i \log[\eta(t)]_i \, , \tag{7}$$

where we use the convention of $0\log(0) \triangleq 0$. The VNE $H(t)$ is dominated by the relatively large eigenvalues, while eigenvalues that are relatively small contribute little. Therefore, it provides a measure of the number of the relatively significant eigenvalues.

The VNE generally decreases as t increases. As mentioned previously, the initial decrease is primarily due to a denoising of the data as less significant eigenvalues (likely corresponding to noise) decrease rapidly to zero. The more significant eigenvalues (likely corresponding to signal) decrease much more slowly. Thus the overall rate of decrease in $H(t)$ is high initially as the data is denoised but then low for larger values of t as the signal is smoothed. As $t \to \infty$, eventually all but the first eigenvalue decrease to zero and so $H(t) \to 0$.

To choose t, we plot $H(t)$ as a function of t as in the first column of Fig. 6. Choosing t from among the values where $H(t)$ is decreasing rapidly generally results in noisy embeddings (second column in Fig. 6). Very large values of t result in an embedding where some of the branches or trajectories are combined together and some of the signal is lost (fourth column in Fig. 6). Good PHATE visualizations can be obtained by choosing t from among the values where the decrease in $H(t)$ is relatively slow, i.e. the set of values soon after the "knee" in the plot of $H(t)$ (third column in Fig. 6 and the PHATE embeddings in Fig. 3). This is the set of values for which much of the noise in the data has been smoothed away, and most of the signal is still intact. The PHATE embedding is fairly robust to the choice of t in this range, as demonstrated in the Methods section. The actual value can be chosen by selecting a t value where the second derivative of $H(t)$ is low.

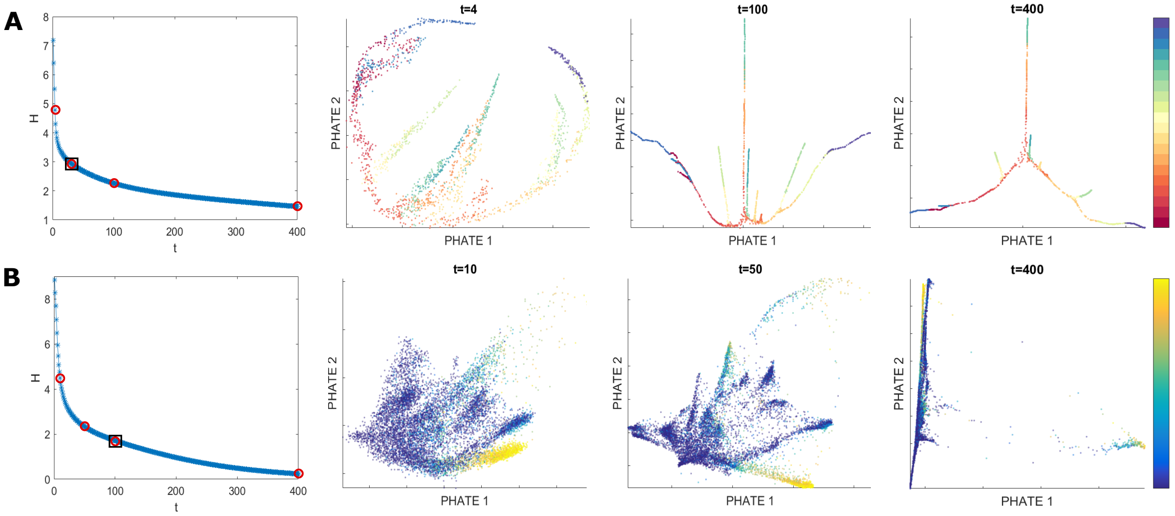

Figure 6: Demonstration of the effect of the scale t on the PHATE embedding for the (**A**) branching data in Fig. 2 and the (**B**) bone marrow mass cytometry data in Fig. 3. The colorings are also the same. The first column shows the VNE $H(t)$ (see Eq. 7) of the diffusion affinities as a function of the time scale t. The other columns give the PHATE embedding with different values of t. The red dots in the first column indicate the values of t chosen for the plots. The red dots surrounded by a black box indicate the chosen value of t for the embeddings in Figs. 2 and 3. Values of t that are too low can give noisy embeddings while very high values of t can result in a loss of information in the embedding. However, the range of t values that give a good embedding is generally quite large.

2.5 Creating Potential Distances

In PHATE, we recover a new type of distance from the transition probabilities of the diffusion operator that we call the *potential distance* by taking the negative log of the transition probabilities. Intuitively, in one-dimensional manifolds (such as branches of a tree), this corresponds to the time it takes to diffuse between two points. Further, in Fig. 7 we show that this transformation has the effect of stabilizing the embedding near the boundaries.

To analyze the constructed heat diffusion process, two possible scenarios can be considered for the origin of the dataset $\mathcal{X}$ and its distribution p, as described in [15, 16]. In the first scenario, the data generation process is modeled as an instantiation of a dynamical system that has reached an equilibrium state independent of the initial conditions. Mathematically, let $U(x)$ be a potential and $w(x)$ be an d-dimensional Brownian motion process. The data distribution is the steady state solution of the of the stochastic differential equation (SDE) $\dot{x} = -\nabla U(x) + \sqrt{2}\dot{w}$, where $\dot{x}$ denotes differentiation of x with respect to time. The time steps of the system are dominated by the forward and backward Fokker-Planck equations. This steady state solution is given by

$$p(x) = \exp(-U(x)),$$

up to normalization in the L^1 norm to form a proper probability distribution.

The distribution of the data in this case is dominated by the potential U that models the underlying structure of the data. As an example, if the data is uniformly distributed on or around a manifold, then this potential is minimal on the manifold itself and increases rapidly when deviating from the manifold. The underlying potential also incorporates data densities that are not uniform. For example, data clusters are represented as local wells or pits in the underlying potential, while progression trajectories and transitions between clusters are represented as rivers or branches in the potential. See [15, 16] for more details.

In the second scenario, the data generation process is not modeled as a dynamical system. Instead, we consider the data in this case as generated by drawing N i.i.d. samples from the probability distribution $p(x)$. We then artificially define the underlying potential of the data as

$$U(x) = -\log(p(x)).$$

The potential U can be used in this scenario since its properties and its relation to the structure of the data are not directly related to the notion of time. Furthermore, in both scenarios, the diffusion-based analysis introduces the notion of diffusion time in order to reveal intrinsic data geometry. Finally, as shown in [15, 16], in both scenarios the Markov process that defines the diffusion geometry converges asymptotically to a diffusion process governed by Fokker-Planck equations with a potential $2U(x)$, whether the original potential is defined naturally or artificially.

Using the same relationship between a potential U and an equilibrium distribution p, we can define a diffusion potential from the stationary distribution p^∞ as $U^\infty = -\log(p^\infty)$. This potential corresponds to data generation using the random walk process defined by P_ε with $t \to \infty$ with random initial conditions. Similarly, if we consider a data generation process using this random walk process with t-steps and a fixed initial condition δ_x, then the generated data is distributed according to p_x^t and the corresponding t-step potential representation of x is $U_{\varepsilon,x}^t = -\log(p_x^t)$.

Given the potential representations U_x^t, $x \in \mathcal{X}$ of the data in $\mathcal{X}$, we define the following potential distance metric as an alternative to the distribution-based diffusion distance:

Definition 1. *The t-step* **potential distance** *is defined as* $\mathfrak{V}^t(x,y) \triangleq \|U_x^t - U_y^t\|_2$, $x, y \in \mathcal{X}$.

The following proposition shows a relation between the two metrics by expressing the potential distance in embedded diffusion map coordinates[1] for fixed-bandwidth Gaussian-based diffusion (i.e., generated by P_ε from Eq. 2):

Proposition 1. *Given a diffusion process defined by a fixed-bandwidth Gaussian kernel, the potential distance from Def 1 can be written as* $\mathfrak{V}^t(x,y) = \left(\sum_{j=1}^{n} \log^2\left(\frac{1+\left\langle\Phi^{t/2}(x),\Phi^{t/2}(x_j)\right\rangle}{1+\left\langle\Phi^{t/2}(y),\Phi^{t/2}(x_j)\right\rangle}\right)\right)^{1/2}$

Proof. According to the spectral theorem, the entries of P_ε^t can be written as

$$[P_\varepsilon^t]_{(x,y)} = \psi_0(y) + \sum_{i=1}^{n-1} \lambda_i^t \phi_i(x)\psi_i(y)$$

since powers of the operator P_ε only affect the eigenvalues, which are taken to the same power, and since the trivial eigenvalue λ_0 is one and the corresponding right eigenvector ϕ_0 only consists of ones. Furthermore, it can be verified that the left and right eigenvectors of P_ε are related by $\psi_i(y) = \phi_i(y)\psi_0(y)$, thus, combined with Eqs. 4 and 6, we get

$$p_{\varepsilon,x}^t(y) = \psi_0(y)\left(1 + \sum_{i=1}^{n-1} \lambda_i^t \phi_i(x)\phi_i(y)\right) = \psi_0(y)\left(1 + \left\langle \Phi_\varepsilon^{t/2}(x), \Phi_\varepsilon^{t/2}(x)\right\rangle\right) .$$

By applying the logarithm to both ends of this equation we express the entries of the potential representation $U_{\varepsilon,x}^t$ as

$$U_{\varepsilon,x}^t(y) = -\log(1 + \left\langle \Phi_\varepsilon^{t/2}(x), \Phi_\varepsilon^{t/2}(y)\right\rangle) - \log(\psi_0(y)) ,$$

and thus for any $j = 1, \ldots, N$,

$$\begin{aligned} \left(U_{\varepsilon,x}^t(x_j) - U_{\varepsilon,y}^t(x_j)\right)^2 &= \big[\log(1 + \left\langle \Phi_\varepsilon^{t/2}(x), \Phi_\varepsilon^{t/2}(x_j)\right\rangle) \\ &- \log(1 + \left\langle \Phi_\varepsilon^{t/2}(y), \Phi_\varepsilon^{t/2}(x_j)\right\rangle)\big]^2 \\ &= \log^2\left(\frac{1 + \left\langle \Phi_\varepsilon^{t/2}(x), \Phi_\varepsilon^{t/2}(x_j)\right\rangle}{1 + \left\langle \Phi_\varepsilon^{t/2}(y), \Phi_\varepsilon^{t/2}(x_j)\right\rangle}\right) , \end{aligned}$$

which yields the result in the proposition. □

[1] Recall the diffusion distance is simply the Euclidean distance in these coordinates

2.6 Diffusion Potential Embedding with Non-Metric MDS

Instead of using diffusion maps coordinates, the potential-based embedding in PHATE is obtained by using the potential distance from Def. 1 as input for distance embedding methods, which find optimal two- or three-dimensional coordinates that approximate the potential distance as an embedded Euclidean distance.

Some common distance embedding methods are known as multidimensional scaling (MDS) methods. Classical MDS [20] takes a distance matrix as input and embeds the data into a lower-dimensional space using eigendecomposition techniques. We apply classical MDS to the potential distances of the data to obtain an initial configuration of the data in low dimension. This configuration is then optimized further using nonmetric MDS as described later in this section. First, we use this initial configuration to demonstrate a crucial advantage of our proposed diffusion potential embedding over diffusion maps.

Consider a simple case of data sampled uniformly on a circle in $\mathbb{R}^2$. Diffusion maps (with suitable density normalization) has been shown to perform well in applications where the data can be modeled intrinsically as being sampled from a circle, e.g., [14, 21, 22]. Indeed, it can be verified in Fig. 7(right) that both the diffusion maps and PHATE embeddings recover the circle up to centering and scaling. However, as a manifold, the circle contains no endpoints, in contrast with the branching structure in many biological datasets. To introduce endpoints, we consider the lower half of the circle in Fig. 7(left). In this case, the diffusion maps embedding suffers from instabilities that generate significantly higher densities near the end points, due to boundary conditions of the diffusion eigenfunctions, which distorts the embedding. The PHATE embedding does not exhibit these instabilities. This demonstrates that the PHATE embedding is more robust than diffusion maps to boundary conditions. Thus, it is better suited for visualizing data with boundary conditions such as those introduced by endpoints as well as branch points, where multiple branches intersect.

While classical MDS is computationally efficient relative to other MDS approaches, it assumes that the input distances directly correspond to low-dimensional Euclidean distances, which may be overly restrictive. Additionally, since we are primarily interested in trajectory visualization, it is not important that the exact distance is preserved between points on two different trajectories.

Nonmetric MDS is an approach that relaxes the assumptions on the distance matrix by allowing the input to be some measure of dissimilarity rather than a distance metric [23–25]. This relaxation is made by optimizing a monotonic relation between the input dissimilarities and the embedded Euclidean distances between the points. This relation is quantified by a *goodness of fit* criterion, which is typically referred to as a **stress** function. Several possible stress functions can be used in nonmetric MDS. The results presented in this paper were obtained by using the

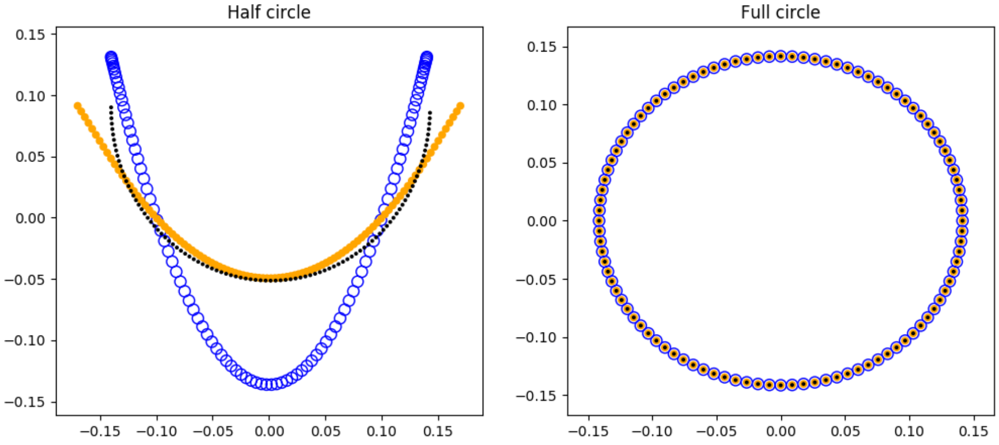

Figure 7: Comparison of Diffusion Maps (blue) and PHATE (orange) embeddings on data (black) from a half circle (left) and a full circle (right). Both the data and the embeddings have been centered about the mean and rescaled by the max Euclidean norm. For the full circle, both embeddings are identical (up to centering & scaling) to the original circle. However, for the half circle, the Diffusion Maps embedding (blue) suffers from instabilities that generate significantly higher densities near the two end points. The PHATE embedding (orange) does not exhibit these instabilities.

popular[2] *Kruskal normalized stress 1* from [25]. Namely, we minimize the following stress:

$$\texttt{Stress}_1(\hat{x}_1,\ldots,\hat{x}_N) = \sqrt{\sum_{i,j}\left(f\left(\mathfrak{D}^t_{(x_i,x_j)}\right) - \|\hat{x}_i - \hat{x}_j\|\right)^2 \Big/ \sum_{i,j}\|\hat{x}_i - \hat{x}_j\|^2}\,. \tag{8}$$

over embedded d'-dimensional coordinates $\hat{x}_i \in \mathbb{R}^{d'}$ of data points in $\mathcal{X}$ and weakly monotone relations[3] $f : \mathbb{R} \to \mathbb{R}$ between potential distances and embedded (Euclidean) distances. This optimization is essentially an isotonic regression problem, which can be solved by suitable standard optimizers (e.g., using gradient descent).

If the stress of the embedded points is zero, then the input data is faithfully represented in the MDS embedding. The stress may be nonzero due to noise or if the embedded dimension d' is too small to represent the data without distortion. Thus, by choosing the number of MDS dimensions to be $d' = 2$ (or $d' = 3$) for visualization purposes, we trade off distortion in exchange for readily visualizable coordinates. However, as mentioned previously, some distortion of the distances/dissimilarities is tolerable in many of our applications since precise dissimilarities between points on two different trajectories are not important as long as the trajectories are

[2] We use the default Matlab `mdscale` implementation.

[3] Technically, the optimization only considers N^2 values of f for distances between points in $\mathcal{X}$.

visually distinguishable. By using non-metric MDS, we find an embedding of the data with the desired dimension for visualization and the minimum amount of distortion as measured by the stress.

2.7 Comparing PHATE to Other Methods

PHATE is primarily a dimensionality reduction method that takes high dimensional raw data and embeds it, via a metric preserving embedding, into low dimensions that naturally show trajectory structure. Thus, we focus our comparisons of PHATE to existing dimensionality reduction methods such as PCA, tSNE and diffusion maps. However, because PHATE can extract trajectory or differentiation structure, we also compare it to tools that find and *render* explicit "differentiation tree structures"; these methods include SPADE [3] and Monocle2 [2].

Finally, we note that several methods exist that find *pseudotime* orderings of cells, such as Wanderlust [4], Wishbone [5], and diffusion pseudotime [26]. These methods focus on finding orderings of cells along branches. These methods can be used alongside PHATE to order parts of the branching progressions. Wanderlust can find single non-branching progressions. Wishbone recognizes a single branch, while diffusion pseudotimes provides potentially multiple branches.

However, pseudotime approaches do not naturally provide a dimensionality reduction method to visualize such structure. Therefore, the resulting cell orderings can be difficult to interpret and verify, especially in the context of the entire data set. In contrast, PHATE reveals the entire branching structure in low dimensions, giving an overall view of progression structure in the data. Thus pseudotime orderings can be visualized and verified with PHATE.

Comparison of PHATE to dimensionality reduction methods: As mentioned previously, Figs. 2 and 3 compare the PHATE embedding to the principal components analysis (PCA), tSNE, and diffusion maps embeddings on four different data sets. For all four datasets, the PHATE visualization is best at distinguishing branches and trajectories within the data. While the diffusion maps embedding does capture some trajectory structure, many of the trajectories are not visible in the visualization. Additionally, the PCA and tSNE embeddings do not emphasize trajectory structure, and the trajectory structures in the data are very difficult, if not impossible, to extract from the PCA and tSNE visualizations. The popular method of principal component analysis (PCA) assumes a linear structure on the data. Since biological data are rarely linear, PCA is not able to optimally reduce non-linear noise along the manifold and reveal progression structure.

Recently, tSNE (t-distributed stochastic neighbor embedding) [1] has become popular for revealing cluster structure or separations in single cell data [27]. However, tSNE tends to shatter trajectories into clusters (Fig. 2), at times artificially. Furthermore, the adaptive kernel used in tSNE for calculating neighborhood probabilities tends to spread out neighbors such that dense clusters occupy proportionally more space in visualization as compared to sparse clusters. Thus, the relative location of data points within the tSNE embedding often does not accurately reflect the relationships between them. Finally, while the diffusion maps embedding does capture some trajectory structure, many of the trajectories are not visible in the visualization, as diffusion maps tend to split trajectories into different orthogonal dimensions instead of showing a unified low-dimensional structure as PHATE shows.

Comparison of PHATE to tree-rendering methods: SPADE [3], Monocle2 [2] and other methods first cluster the data and then render progression as connections between clusters. SPADE finds a minimal spanning tree that fits to the clusters, and Monocle2 finds a graph to fit to the clusters. Clustering methods tend to make less restrictive assumptions on the structure of the data compared to PCA. However, clustering methods assume that the underlying data can be partitioned into discrete separate regions. In reality, biological data are often continuous, and the apparent cluster structure given by clustering methods is only a result of non-uniform density and finite sampling of the continuous underlying state space. Further, these methods tend to be unstable, producing different trees and different numbers of branches each time that they are run, as shown in Figs. 8B and 8C. Thus it is difficult to determine the right tree to fit to the data. Further, several spurious branches seem to arise in both settings. In contrast, for the same set of parameters, PHATE produces the same results with each run as it is not based on a tree or graph-fitting paradigm.

2.8 PHATE Overlays

In this section, we describe some methods for automatically extracting information from the PHATE embedding. We first describe techniques for identifying branch points using local intrinsic dimensionality and end points using eigencentrality and diffusion map extrema within the embedding. From these points, trajectories can be extracted for analysis.

2.8.1 Branch Point Identification with Local Intrinsic Dimensionality

A PHATE embedding consists of trajectories and branching points. Trajectories are paths of progression along which cells vary smoothly in particular dimensions. Branch points are decision points where cells sharply veer towards one of a small number of fates, and contain switch-like decisions. For instance, there is a split between CD4+ cells and CD8+ cells in Fig. 11A, where CD4 is turned off in one branch and CD8 in another. These represent distinct mutually exclusive paths of progression.

We use the concept of local intrinsic dimensionality for identifying these types of branch points. In biological data, often many variables for each datapoint are measured. The total number of variables measured for each data point is the extrinsic dimension of the data. However, generally many dependencies and redundancies exist between these variables. Thus, the total number of (potentially transformed) variables required to accurately represent the data is less than the extrinsic dimension. This number is known as the intrinsic dimension of the data.

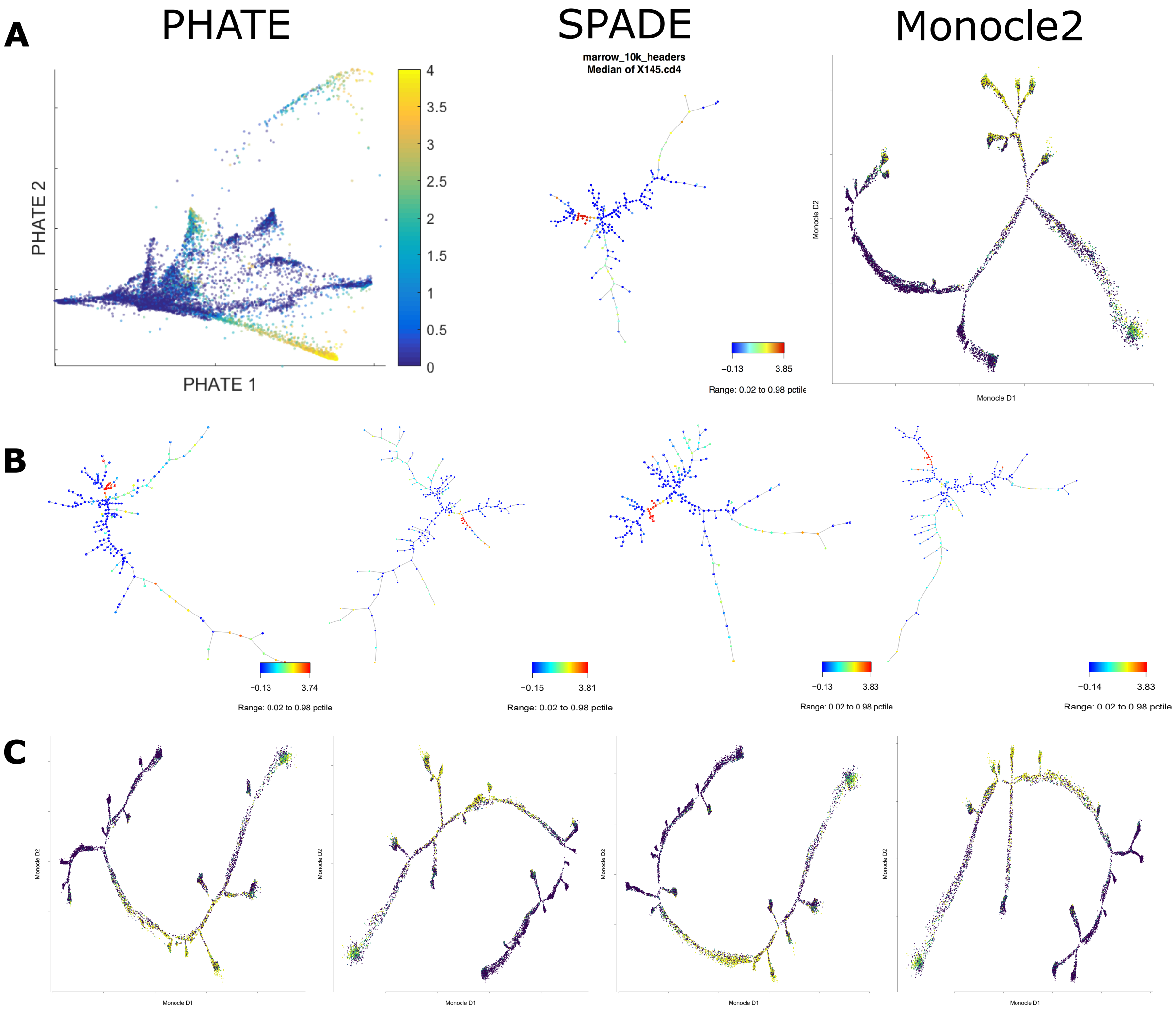

Figure 8: (**A**) Comparison of PHATE to SPADE and Monocle2 on the bone marrow mass cytometry data set [9] colored by CD4 expression levels. (**B**) Multiple runs of SPADE on the same data set colored by CD4 expression levels. (**C**) Multiple runs of Monocle2 on the same data set colored by CD4 expression levels. Some of the results from the different runs for both SPADE and Monocle2 vary significantly from each other suggesting that they are sensitive to randomization. In contrast, given the same parameters, PHATE produces the same results with each run.

Intrinsic dimension can also be understood in terms of manifolds. If the dependencies between variables are smooth, then the data can be modeled as lying on a manifold inside the full space with intrinsic dimension less than the extrinsic dimension of the full space. For many complicated datasets, the underlying manifold, and potentially the intrinsic dimension, of the data may also vary *locally* [28, 29]. For a toy example of data with varying local intrinsic dimension, see Fig. 9A. In this figure, the red data points can be modeled as lying on a manifold with intrinsic dimension equal to one (a circle) while the black data points can be modeled with a manifold with intrinsic dimension equal to two (a plane).

Intuitively, points on branches lie on manifolds with low intrinsic dimension. Branch points are regions where two or more branches originate or intersect. Thus, branch points are locations where several directions of progression merge into a cluster of data points. This cluster lies on a manifold with higher intrinsic dimensionality than the branches. This suggests that local intrinsic dimensionality estimation techniques may be used to detect branching zones.

There are many different methods for estimating local intrinsic dimension. We use the method given in [28], which uses a local version of the k-nn graph approach derived in [30] combined with neighborhood smoothing for variance control as follows. Let $\mathbf{Z}_n = \{\mathbf{z}_1, \dots, \mathbf{z}_n\}$ be a set of independent and identically distributed random vectors with values in a compact subset of $\mathbb{R}^d$. Let $\mathcal{N}_{k,j}$ be the k nearest neighbors of $\mathbf{z}_j$; i.e. $\mathcal{N}_{k,j} = \{\mathbf{z} \in \mathbf{Z}_n \setminus \{\mathbf{z}_j\} : ||\mathbf{z}-\mathbf{z}_j|| \leq \epsilon_k(\mathbf{z}_j)\}$. The k-nn graph is formed by assigning edges between a point in $\mathbf{Z}_n$ and its k-nearest neighbors. The power-weighted total edge length of the k-nn graph is related to the intrinsic dimension of the data and is defined as

$$\mathbf{L}_{\gamma,k}(\mathbf{Z}_n) = \sum_{i=1}^{n} \sum_{\mathbf{z}\in\mathcal{N}_{k,i}} ||\mathbf{z} - \mathbf{z}_i||^{\gamma}, \tag{9}$$

where $\gamma > 0$ is a power weighting constant. Let m be the global intrinsic dimension of all the data points in $\mathbf{Z}_n$. It can be shown that for large n,

$$\mathbf{L}_{\gamma,k}(\mathbf{Z}_n) = n^{\beta(m)}c + \epsilon_n, \tag{10}$$

where $\beta(m) = (m-\gamma)/m$, ϵ_n is an error term that decreases to 0 as $n \to \infty$, and c is a constant with respect to $\beta(m)$ [30]. A global intrinsic dimension estimator $\hat{\mathbf{m}}$ can be defined based on this relationship using non-linear least squares regression over different values of n [28, 30].

A local estimator of intrinsic dimension $\tilde{\mathbf{m}}(i)$ at a point $\mathbf{z}_i$ can be defined by running the above procedure in a smaller neighborhood about $\mathbf{z}_i$. This approach is demonstrated in Fig. 9A, where a k-nn graph is grown locally at each point in the data. However, this estimator can have high variance within a neighborhood. To reduce this variance, majority voting within a neighborhood of $\mathbf{z}_i$ can be performed:

$$\hat{\mathbf{m}}(i) = \arg\max_{\ell} \sum_{\mathbf{z}_j\in\mathcal{N}_{k,i}} \mathbb{1}(\tilde{m}(j) = \ell), \tag{11}$$

where $\mathbb{1}(\cdot)$ is the indicator function [28].

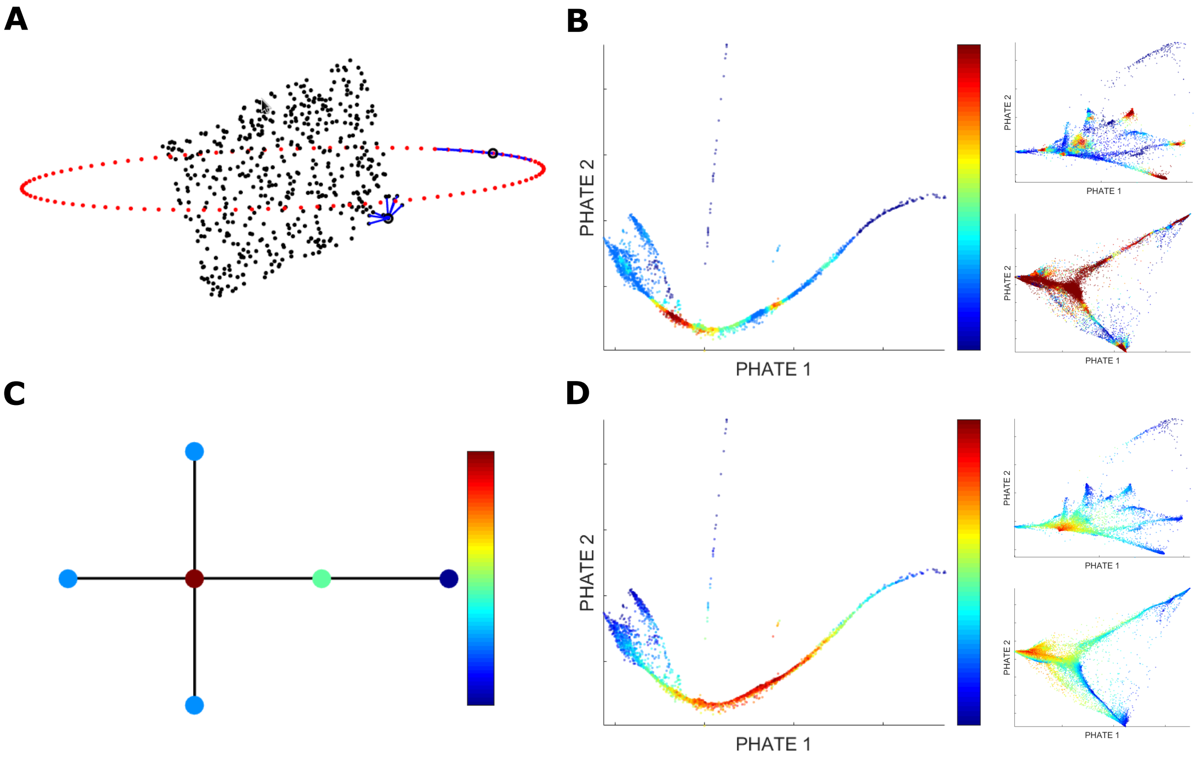

Figure 9: (**A**) A toy example of data with varying local intrinsic dimension. The red data points can be modeled as lying on a manifold with intrinsic dimension equal to one (a circle) while the black data points can be modeled with a manifold with intrinsic dimension equal to two (a plane). Local intrinsic dimension can be estimated by growing a k-nn graph locally at each data point due to the relationship between the k-nn graph growth rate and intrinsic dimension. (**B**) The PHATE embedding of the bone marrow scRNAseq dataset shown in Fig. 3A colored by estimated local intrinsic dimensionality for higher dimensional ($d' = 10$) PHATE embeddings. The estimated local intrinsic dimension is higher at the branch points compared to the branches. The results are also shown for the mass cytometry datasets. (**C**) A small graph with nodes colored by eigenvector centrality. The node that is most connected has the highest centrality and the centrality of the other nodes depends on their proximity to the most connected node as well as their connectivity. (**D**) The PHATE embedding of the bone marrow scRNAseq dataset shown in Fig. 3A colored by eigenvector centrality calculated from the affinity matrix $K_{k,\alpha}$. The endpoints of the branches have lower centrality than other points. The results are also shown for the mass cytometry datasets.

Figure 9B shows the estimated local intrinsic dimensionality for higher dimensional ($d' = 10$) PHATE embeddings of the bone marrow scRNAseq dataset shown in Fig. 3A. The estimated local intrinsic dimension is higher at the branch points compared to the branches.

2.8.2 Endpoint Identification with Eigencentrality and Diffusion Map Extrema

In addition to branch points, we also identify end points in the PHATE embedding. These points correspond to end-states of differentiation processes. We use two features of the data to accomplish this: eigenvector centrality and diffusion maps extrema.

The centrality of a graph is a measure of the relative influence of a node within a graph. Nodes with higher centrality have more influence than nodes with lower centrality. Eigenvector centrality of a graph is a measure of graph centrality. It is defined as the eigenvector corresponding to the largest eigenvalue of the corresponding adjacency matrix [31]. The adjacency matrix we use is the kernel matrix $K_{k,\alpha}$. Points located at the ends of branches in the PHATE embedding have less influence on the graph. Thus the eigenvector centrality of these points should be relatively lower (see Fig. 9C). Figure 9D shows the computed eigenvector centrality of the kernel matrix $K_{k,\alpha}$ derived from the bone marrow scRNAseq dataset. Indeed we find that the end points of the branches generally have lower eigenvector centrality.

While choosing points with low eigenvector centrality successfully identifies some end points of branches, some endpoints may have relatively higher centrality due to their proximity to regions with high centrality. For example, the endpoint of the left-most branch in the PHATE embedding of the bone marrow mass cytometry data set is closer to the most central region of the data than the endpoints of the branches on the right (see Fig. 9D). Thus the eigenvector centrality of the left branch endpoint is relatively higher than the eigenvector centrality of the right branch endpoints. Therefore, a global threshold on the eigenvector centrality that is high enough to include this left branch endpoint would also include many other points on other branches that are not endpoints.

We automatically detect such endpoints by using the extrema of the diffusion maps embedding. The diffusion maps embedding tends to map the endpoints of branches into the minimum and maximum values of different dimensions, including the endpoints that have relatively higher eigenvector centrality (see the DM embeddings in Fig. 3). Thus we can identify many of these points by choosing the points with the minimum and maximum values in the first few diffusion maps dimensions.

2.8.3 Branch Point and Endpoint Reduction

After identifying branch points and end points, it becomes easier to identify the segments or trajectories of smooth progression in the data. However, since eigencentrality and local intrinsic dimension vary smoothly, they tend to select regions in the embeddings rather than particular points. Therefore, we use a simplified version of the *shake-and-bake* algorithm from [32] to reduce the number of branch points and endpoints to a smaller set of representative points;

namely, ones that correspond to unique decisions and end-states in the data. Algorithm 2 details the steps of this procedure, based on a proximity threshold that determines the smallest possible distance desired between the representative branch points and/or endpoints. The candidate branch points are given in coordinates obtained from applying MDS to potential distances, but with an embedded dimension higher than two. The presented results in this paper were obtained with this dimension set to 10.

Algorithm 2: Shake-and-Bake branch & end point reduction

```
Input: Branch point candidates $X = \{x_1, x_2, \ldots\}$, proximity threshold $h$
Output: Branch points $R$
$D \leftarrow$ compute pairwise distance matrix from $X$
$\mathcal{I} \leftarrow$ random permutation of indices over $X$
$R \leftarrow \emptyset$
for $j \in \mathcal{I}$ {pop the next index based on the permuted order} do
    Set neighborhood: $\mathcal{N}_j \leftarrow \{i \in \mathcal{I} : \|x_i - x_j\| \leq h\}$
    $r_j \leftarrow$ centroid of points in $\{x_i \in X : i \in \mathcal{N}_j\}$
    Add centroids: $R \leftarrow R \cup \{r_j\}$
    Remove neighbors: $\mathcal{I} \leftarrow \mathcal{I} \setminus \mathcal{N}_j$ {maintain permuted order of remaining indices}
end for
```

3 PHATE reveals insights into biological differentiation processes

PHATE reveals branching differentiation structures in biological data. In this section, we show the insights gained through the PHATE visualization, which is able to reveal paths of progression, decision or branch points, and end-states within the various biological datasets used in Fig. 3, new embryoid body data, SNP data, and microbiome data.

3.1 PHATE Trajectories Have Biological Meaning

We show that the identifiable trajectories in the PHATE embedding have biological meaning that can be discerned from the expression and mutual information of genes along the trajectories. Figures 10 and 11 show the results for the bone marrow scRNAseq [8], bone marrow CyTOF [9], and IPSC CyTOF [10] datasets. For each of these datasets, we manually selected trajectories between the representative branch points and endpoints (explained in Section 2.8.3). We then ordered the cells within each trajectory by projecting the cells onto the line corresponding to the branch. Ordering is generally from left to right. We note that we could also order these points based on pseudotime ordering software such as those in [4], [5] or [26].

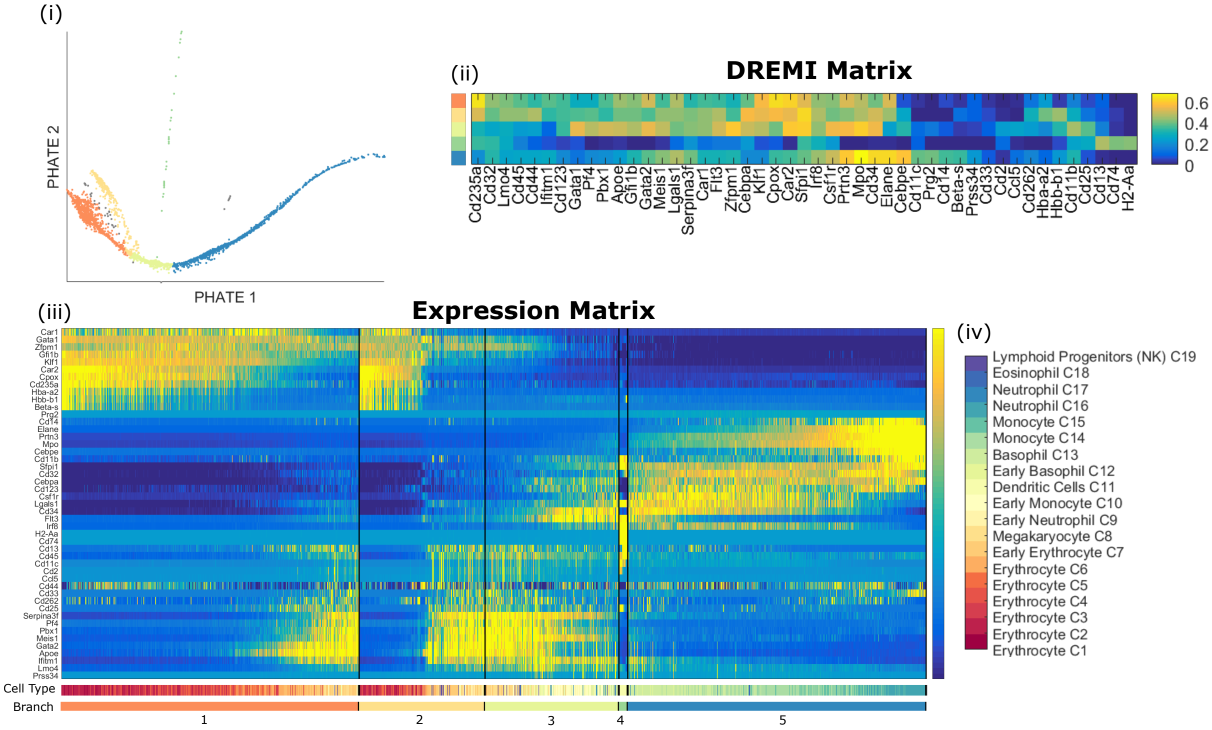

Figure 10: Analysis of branches on the PHATE embedding for the bone marrow scRNAseq dataset from Fig. 3. (i) The PHATE embedding with identified branches. (ii) DREMI scores [33] between gene expression levels and cell order within each branch. Cell ordering is from the leftmost to the rightmost endpoint of each branch. MAGIC [34] is applied to the scRNAseq data first before computing DREMI to impute missing values in the data. (iii) Expression level for each cell ordered by branch and ordering within the branch. MAGIC is applied first with the same kernel used for PHATE and scale $t = 4$. Expression levels are then z-scored for each gene. (iv) Legend for the cell types identified in [8]. A colorbar is also given below the expression matrix in (iii) that identifies each cell's type.

Figures 10 and 11 show the PHATE embedding for the three datasets with the trajectories identified by color along with gene expression matrices that show the expression level of each cell along the trajectory. These matrices show the expression of genes along the identified branches. Ubiquitously expressed genes along a trajectory can allow us to identify cells of the trajectory. Additionally, we show DREMI matrices that show the mutual information between the cell order within each branch and selected protein markers to show which genes change along the branch to form the progression. DREMI is a conditional-density resampled mutual information, that takes off sampling biases to reveal shape-agnostic relationships between two variables [33]. As applied here it shows markers that change along a trajectory.

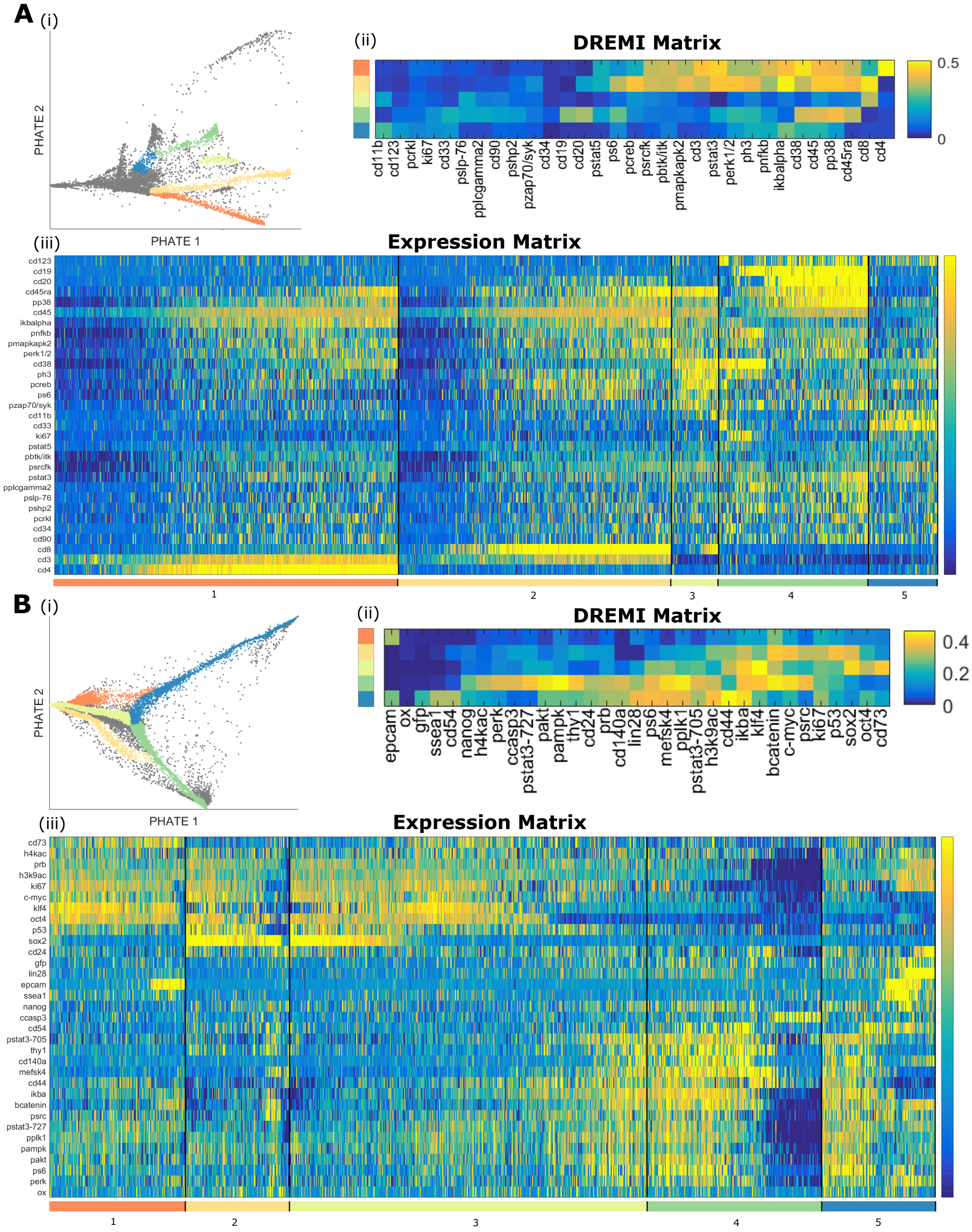

Figure 11: A similar branch analysis as in Fig. 10 applied to the (**A**) bone marrow mass cytometry dataset and the (**B**) iPSC mass cytometry data set from Fig. 3. MAGIC is not applied to this data.

Bone Marrow scRNAseq Data Figure 10 shows the color-coded embedding, DREMI matrix, and gene expression matrix for single-cell RNA-sequencing data from mouse bone marrow. This data is enriched for myeloid and erythroid lineages and was organized into clusters in [8], which are provided in Fig. 10(iv). Here, we show that PHATE reveals a continuous progression structure instead of cluster structure and illustrates the connections between clusters. The PHATE embedding shows a continuous progression from progenitor cell types (shown in light green in the “Cell Type” color bar below the expression matrix) to erythroid lineages (in red) towards the left and myeloid lineages towards the right (in cooler colors). The expression matrix shows increasing expression of erythroid markers in the leftmost branches (branches 1 and 2) such as hemoglobin subunits Hba-a2 and Hbb-b1 as well as heme synthesis pathway enzyme Cpox as the lineage progresses to the left. Towards the right in branch 5, we see an enrichment for myeloid markers, including CD14 and Elane, which are neutrophil markers. In addition, PHATE splits the erythrocytes into two branches not distinguished by the authors of [8]. These branches show differential expression of several genes. Branch 1 is more highly expressed in Gata1 and Gfi1B, both of which are involved in erythrocyte maturation. Branch 2 is more highly expressed in Zfpm1 which is involved in erythroid and megakaryocytic cell differentiation. Given these differential expression levels, it is likely that branch 1 contains erythrocytes that are still maturing while branch 2 contains erythrocytes that are fairly mature [35–41]. In addition, the branches towards the right have high mutual information with CD235a, which is an erythroid marker that progressively increases in those lineages, and also with CD34, which progressively decreases in that lineage.

We note that due to the lack of common myeloid progenitors in this sample, a gap is expected in the PHATE embedding between the monocytes and megakaryocyte lineage since PHATE does not artificially connect separable data clusters (see Fig. 19). However, we note that both the tSNE and PCA embeddings of this data in Fig. 3 also lack a gap between these trajectories. Given that tSNE in particular is designed to separate clusters, this lack of separation is likely due to low cell number and depth of measurements in the data.

Bone Marrow Mass Cytometry Data Figure 11A shows an early CyTOF dataset from a human bone marrow. Branches in this dataset show both developing lineages (B cells, immature neutrophils) as well as developed T cell subtypes, also identified in [9]. Here, we see that the branches can be identified as CD4+ helper T Cells in Orange, CD8+ cytotoxic T cells in yellow, B cells in green as well as developing leukocytes (possibly immature neutrophils) in blue. Additionally, the light green branch appears to be natural killer cells as identified in [9], which express CD38 and some of which also express CD8.

iPSC Mass Cytometry Data Figure 11B is a mass cytometry dataset from [10] that shows cellular reprogramming with Oct4 GFP from mouse embryonic fibroblasts (MEFs) to induced pluripotent stem cells (iPSCs) at the single-cell resolution. The protein markers measure pluripotency, differentiation, cell-cycle and signaling status. The cellular embedding (with combined timepoints) by PHATE shows a unified embedding that contains five main branches, each corresponding to biology identified in [10]. The light green cells represent early reprogramming intermediates with the correct set of reprogramming factors Sox2+Oct4+Klf4+Nanog+ without CD73+ or CD104+. Out of the light-green stem emerges two branches. The blue branch on top shows the successfully reprogramming ESC-like lineages expressing markers such as Nanog, Oct4, Lin28 and Ssea1, and Epcam that are associated with transition to pluripotency [42]. The green branch shows a lineage that is refractory to reprogramming, does not express pluripotency markers and is referred to as still “mesoderm-like” in [10].

Then, the side orange branch represents an intermediate, partially reprogrammed state also containing Oct4+Klf4+CD73+ but is not yet expressing pluripotency markers like Nanog or Lin28. However, the PHATE embedding indicates that as Epcam, which is known to promote reprogramming generally [43], increases along this branch (as evidenced by its high DREMI score against the branch). It joins into the blue branch at a later stage, showing perhaps an alternative path or timing of reprogramming. Finally, the yellow branch shows a lineage that has failed to reprogram successfully perhaps due to the wrong stoichiometry of the reprogramming factors [44]. Of note, this lineage does not contain Klf4+ which is an essential reprogramming factor.

Additionally, the PHATE embedding shows a decrease in p53 expression in precursor branches (light green and yellow) indicating that these cells are released from cell cycle arrest induced by initial reprogramming factor over expression [45]. However, along the green refractory branch we see an increase in cleaved-caspase3, potentially indicating that the failure to reprogram correctly initiates apoptosis in these cells [10].

3.2 PHATE on Embryoid Body Data

Embryonic stem cell (ESC) differentiation is a multi-step process that begins with the induction of primary germ layers –ectoderm, endoderm and mesoderm. *In vitro*, the induction of primary germ layers occurs spontaneously when ESCs are grown as three-dimentional aggregates called embryoid bodies (EB) in the absence of self-renewing signals. EB differentiation closely resembles the embryo development *in vivo* and has been successfully used to produce multiple cell types, including various types of neurons, astrocytes and oligodendrocytes [46–49], hematopoietic, endothelial and muscle cells [50–58], hepatocytes and pancreatic cells [59, 60], as well as germ cells [61, 62]. However, this process is inefficient. The molecular pathways regulating germ layer development are largely unknown. It remains unclear whether in vitro-derived cells represent genuine functional cell types. A deeper and more systematic understanding of human ESCs differentiation is necessary to overcome these challenges. Here, we begin developing such an understanding, using single-cell technology combined with PHATE to elucidate paths of differentiation and gene-gene interactions that underlie differentiation.

We generated new scRNAseq data from a 27-day long EB differentiation timecourse. To comprehensively sample developmental transitions over time, we collected EBs with 3 day intervals, and then combined them in pairs – day 0 with day 3, day 6 with day 9, and so on. EBs were dissociated into single cells, FACS-sorted to remove doublets and dead cells, and

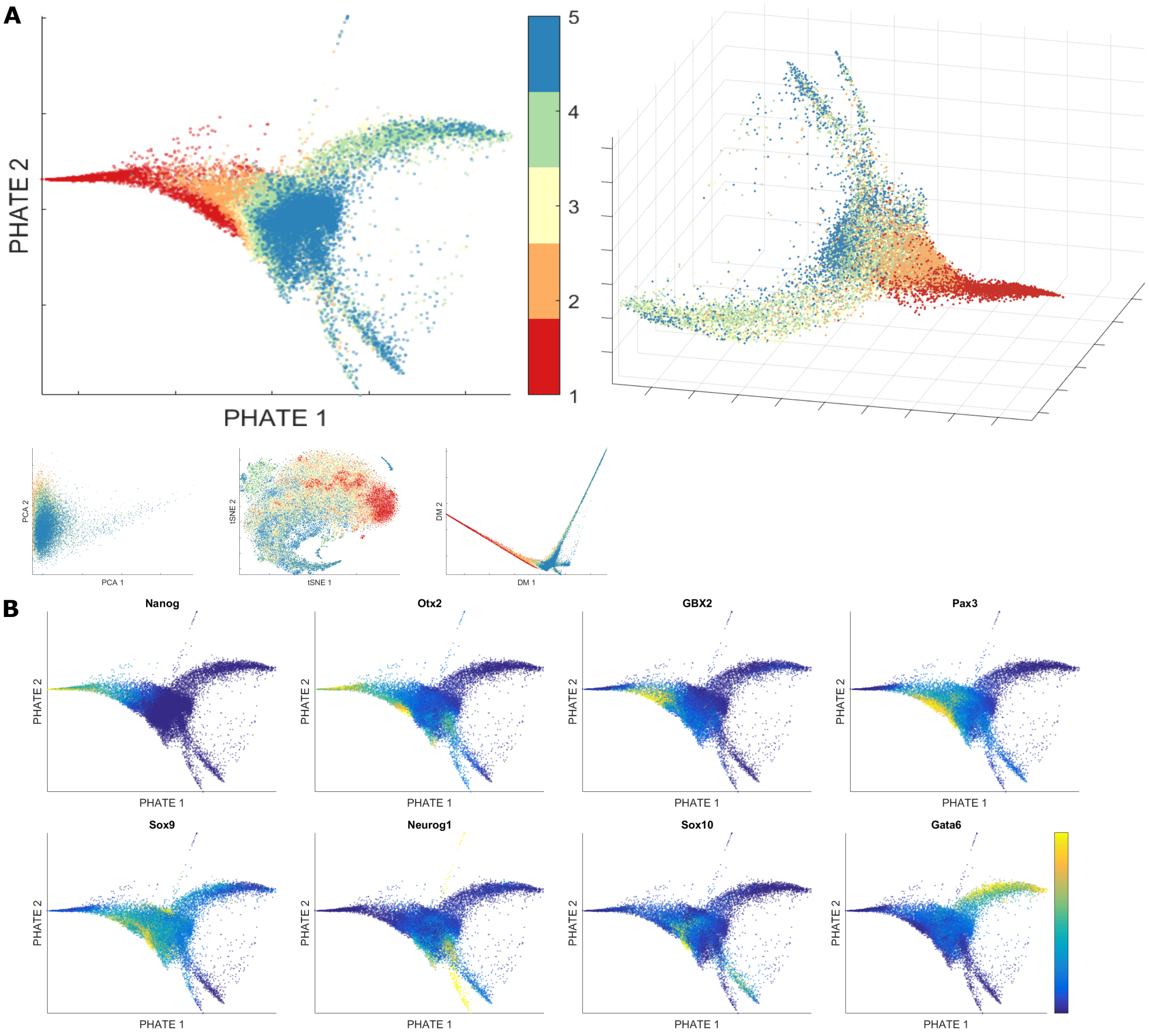

Figure 12: Analysis of new hESC scRNAseq data. (**A**) The PHATE embedding in two and three dimensions compared to PCA, tSNE, and DM on the hESC data. The scale for DM and PHATE is $t = 25$. Cells are colored by sample. The PHATE embedding shows a clear branching structure that is correlated with the samples. (**B**) The PHATE embedding colored by z-scored expression levels of various markers. MAGIC is applied first using the same kernel as for PHATE and scale parameter $t = 4$.

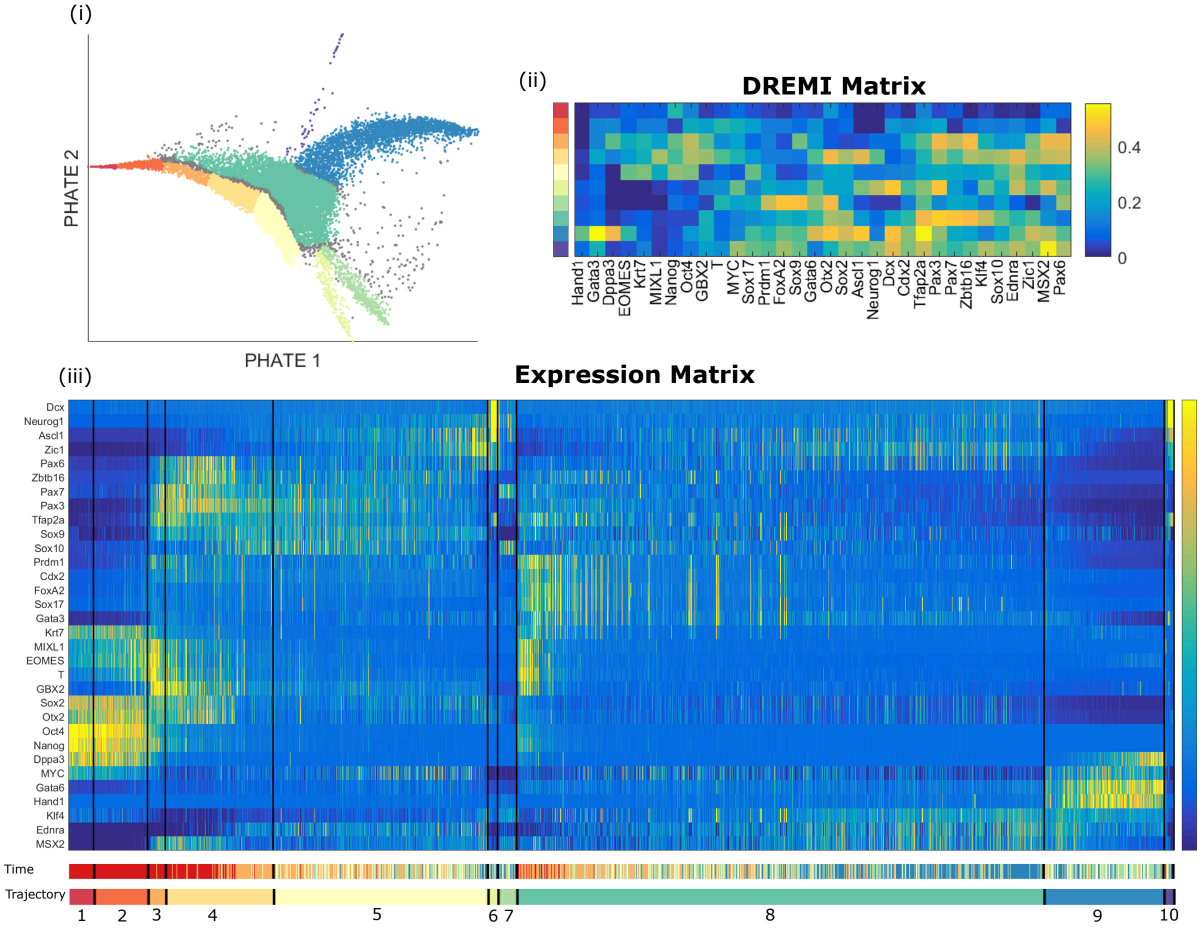

Figure 13: Branch analysis of the new hESC scRNAseq data. Parts (i) through (iii) are constructed in the same manner as in Fig. 10. The time colorbar below the expression matrix corresponds to the sample color in Fig. 12.

processed on a 10x genomics instrument resulting in approximately 31,000 cells equally distributed over the timecourse. Figure 12A shows PHATE applied to this EB data compared to PCA, tSNE, and DM. All embeddings are colored by sample.

The PHATE embedding shows a clear branching structure that is correlated with the samples. Using the PHATE embedding, we can identify several different stages and lineages in the data. Figure 13(i) shows the PHATE embedding colored by trajectories or clusters of cells identified using the process described previously as well as the markers specifically expressed in those trajectories. Figure 13(iii) shows the corresponding expression matrix of selected genes ordered in the same manner as in Fig. 10. From this matrix, we see that trajectories 2-7 are associated with the neural crest differentiation. Along this trajectory, the ES cell genes Nanog, Oct4, and Sox2 are sharply downregulated followed by induction of epiblast marker Otx2 and then neuroectoderm/early neural crest markers Pax6, Zbtb16, Gbx2, Pax3a, and Pax7 in trajectories

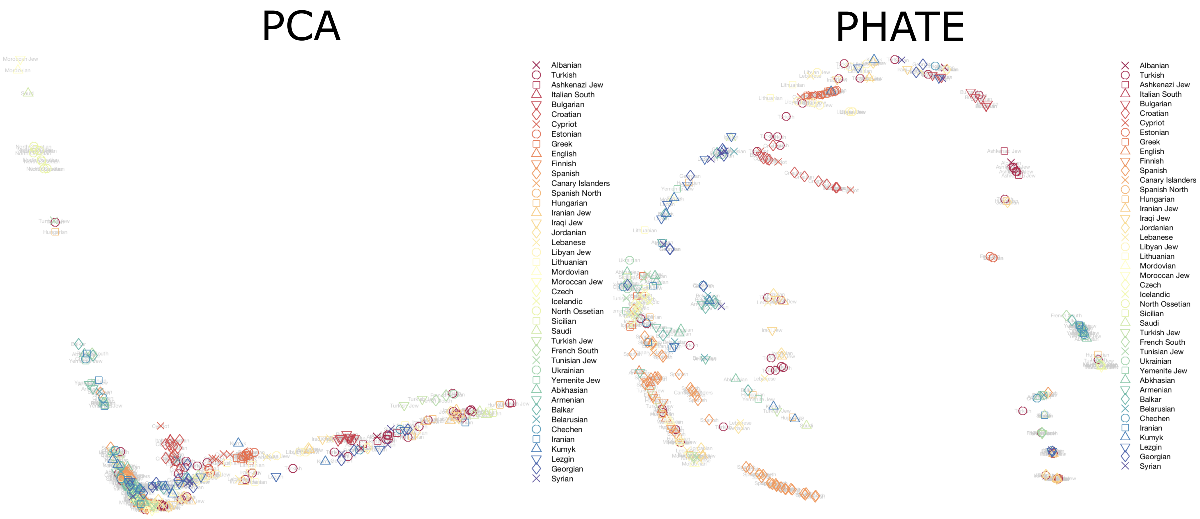

Figure 14: PCA and PHATE embeddings of the Human Origins dataset showing genotyped present day humans from 203 populations.

3 and 4. These early progenitors further resolve into neuronal and late neural crest lineages with characteristic markers expressed in each branch. The neural crest trajectories 5 and 6 express Pax7, Sox9, and Sox10 while neural progenitor trajectories 7 and 10 express Ascl1, Neurog1, and Dcx. Interestingly, differentiation intermediates in trajectory 3 express genes associated with both mesendoderm (Eomes, T, Mixl1) and early neural crest (Pax3, Pax7, Tfap2a), indicating that the separation of these germ layers occurs at this timepoint. Indeed, mesendoderm markers continue to be expressed in trajectory 8 and are followed by a wave of the definitive endoderm markers Foxa2 and Sox17. Trajectory 9 represents cardiac-progenitor-like cells that express Gata6 and Hand1. Thus the PHATE embedding can successfully resolve germ layers during in vitro differentiation of human ESCs.

3.3 PHATE on SNP Data

In Section 2.1 we delineated two features of data that PHATE takes advantage of, namely: 1. data with development that occurs incrementally, as an aggregation of many small modifications, and 2. data with a limited number of possible outcomes from each incremental modification. However, single-cell data is not the only type of biological data that has this type of structure. Genetic data such as single-nucleotide polymorphism data on populations can have such structure too. Individuals, like cells can be slightly modified from each other, and populations as a whole can diverge in a limited number of ways. To demonstrate that PHATE emphasizes trajectory structure in this type of data, we examined a dataset containing 2345 present-day humans from 203 populations genotyped at 594,924 autosomal single nucleotide polymorphisms (SNPs) with the Human Origins array [63].

We used the Eigensoft package [64] to extract 100 PCA components from the SNP array data. As with single-cell RNA sequencing data, we computed the distance and affinity matrices using these PCA components. Figure 14 shows that PHATE is able to reveal geographic population structure much more clearly than PCA alone.

PCA tends to crowd populations together into two linear branches, without clearly distinguishing between population groups or showing population divergence. Vastly divergent populations, such as Near-Eastern populations (e.g., Saudi and Iraqi) are mixed with Eastern European populations (e.g., Ukranian and Bulgarian) and Northern European populations (e.g., Finnish and Estonian). These populations are crowded along the bottom axis of the PCA embedding. By contrast, the PHATE embedding shows clear population structures, such as the near eastern Jewish populations near the bottom (Iranian and Iraqi Jews, Jordanians), with further branches showing progression within the same population, such as the Jordanian population in orange diamond. Further, PHATE shows a global structure that mimics geography, with European populations generally towards the top and Near Eastern populations towards the bottom. Thus PHATE shows that the occurrence and structure of these SNPs follows a progression based on geography and population divergence. Further, as compared to PCA, the plot is highly denoised, as the very high dimensional SNP structure lies in lower dimensional manifolds that are captured by the Markov affinity matrix and denoised via the diffusion process within PHATE.
.

3.4 PHATE on Microbiome Data

Recently there have been various studies of bacterial species abundance in the human intestinal tract, saliva, vagina and other membranes as measured by 16S ribosomal-RNA-encoding gene. It is hypothesized that the bacterial composition of the intestinal tract can affect a wide range of health and metabolic issues such as body mass, autoimmunity, glycemic index, etc. However, generally this data has only been analyzed by clustering and principal component analysis.

A prominent study reported that there were three distinct clusters designated as “enterotypes” identifiable by the variation in the levels of one of three bacterial genera [65]: Bacteroides (enterotype 1), Prevotella (enterotype 2) and Ruminococcus (enterotype 3). We study these enterotypes on the American Gut Dataset [66], a public repository of over 6500 individuals whose tissues have been sequenced by 16s sequencing. Figure 15A shows 9660 samples embedded with PHATE. First we note that PCA (Fig. 15A left) results in an undifferentiated cloud with two density centers corresponding to fecal samples on the right and oral/skin samples on the left. In contrast, PHATE shows branching structures with 4 branches emanating from a point of origin for fecal sample, and additional structures on the right that differentiates between skin samples, which form their own progression, and oral samples, which again result in several branches.

Figure 15B shows the PHATE embedding colored by two genera (bacteroides and prevotella) and a phylum (actinobacteria) of bacteria on the same 9660 samples as in Fig. 15A. These two figures show that the Bacteroides genus of bacteria is almost exclusively found in the fecal samples. The Prevotella genus of bacteria is found in certain stool and oral samples while the Actinobacteria phylum is primarily found in the oral and skin samples. This is consistent with the work in [67] which showed that different genera and phyla of bacteria are prevalent in the different body sites.

Upon "zooming in" to the 8596 fecal samples in Fig. 15C, we see 4 major branches, instead of the three enterotypes [65], with highly expressed Firmicutes, Prevotella, Bacteroides and Verrucomicrobia respectively. Furthermore, the Fermicutes/Bacteroides branches seem to form a smooth continuum with samples falling into various parts of a triangular simplex shape. This shows that individuals can exist as mixed phenotypes between archetypal bacterial states as well as in a continuum with more or less prevalence for each of these states. This could have implications in metabolism and disease of individuals. For instance, it has been noted that individuals with a primarily carbohydrate-based diet have predominantly Prevotella in their gut while individuals who consume more animal fat and proteins have more Bacteroides [68]. These types of causal dietary associations would be easier to extract via correlation with trajectories rather than simple expression analysis along clusters.

4 Conclusion

Modern high-dimensional, high-throughput datasets are difficult for biologists to interpret. Therefore, visualization and data-exploration tools are key to understanding and extracting meaningful structure in biological data and then generating experimental hypotheses. A key observation we make here is that biological datasets have predominant progression structures that most visualization methods do not naturally emphasize. The PHATE method presented in this paper provides a complete embedding and visualization of such branching progression structures in two dimensions, while simultaneously denoising the data. It is based on metric embedding of a novel diffusion potential distance, which is recovered from the Markov data-driven diffusion operator. This metric enables PHATE to express data *trajectories* in low-dimensional coordinates, in contrast to other methods, such as PCA, diffusion maps, or tSNE.

We showed that PHATE can be colored by gene expression, local intrinsic dimensionality, and eigencentrality, which reveal progressions of gene expression (or other biological variables) along branching trajectories, identify branch points where lineages diverge (or converge), and identify end-state cell types in the embedding, respectively. We further demonstrated that biologically meaningful progressions in several single-cell datasets can be showcased by PHATE. These include, for example, cells developing in the bone marrow measured by mass cytometry and single-cell RNA-sequencing, embryoid body differentiation measured with single-cell RNA-sequencing, and induced pluripotent stem cell programming as measured by CyTOF. Finally, our results showed that non-single cell datasets, such as population genetics SNP data and gut microbiome data, also have progression structures where individuals vary slightly from each other in a way that can be modeled as forming latent branching progressions.

Our results indicate that PHATE is able to provide meaningful biological insights from developmental data, including the ability to derive what drives biological progressions. For in-

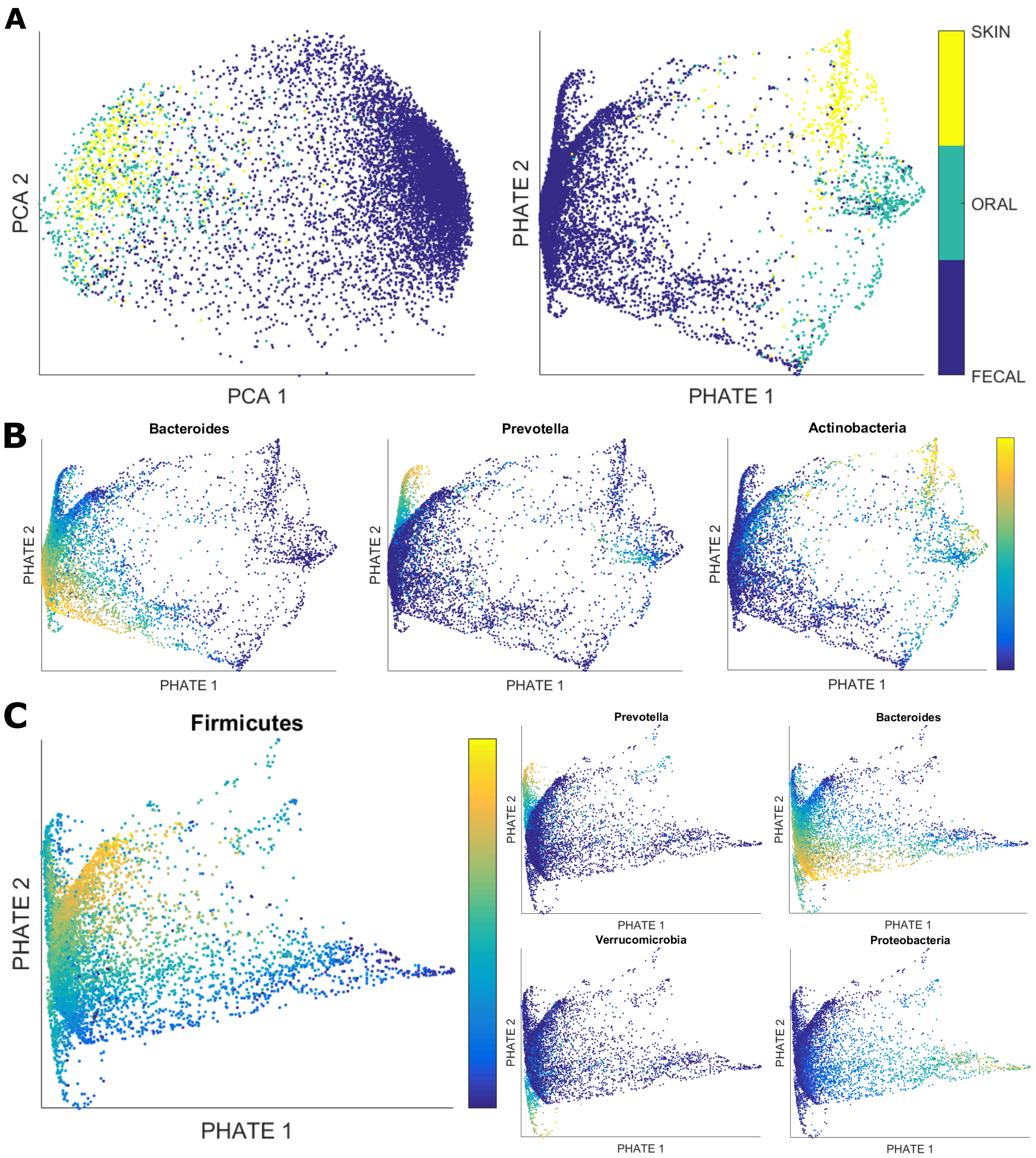

Figure 15: Analysis of data from the American Gut project. (**A**) PCA and PHATE embeddings colored by body site. PHATE shows multiple branches that are not visible in the PCA embedding. (**B**) The PHATE embedding colored by 2 genera (bacteroides and prevotella) and a phylum (actinobacteria) of bacteria. (**C**) The PHATE embedding of only the fecal samples colored by various genera (bacteroides and prevotella) and phyla (firmicutes, verrucomicrobia, and proteobacteria) of bacteria. Each PHATE branch is associated with one of these bacteria groups.

stance, we showed that computing the mutual information-based score DREMI along trajectories can lead to insights on what drives progressions. Furthermore, PHATE shows the complete branching structure within the data. Therefore, paths of progression, how they intersect and alternative shortcut paths (such as the alternative paths shown in IPSC) are visually evident in the PHATE graphical visualization. This can suggest alternative paths of reprogramming for IPSCs or alternative developmental branches not usually seen in experiments. Additionally, branch-points or decision points are easy to decipher on a PHATE embedding. These points can be examined to learn cellular logic, i.e., which genes create the split after a branch point and which genes switch on and off.

Future work will involve using the presented approach for further analyses, as well as experimental validation of gained insights on, for example, embryoid body data. Additionally, the scalability of PHATE will be enhanced by applying optimization and numerical techniques, such as sampling, dictionary learning, out of sample extension, data fusion, and deep learning. In particular, we will explore scalable alternatives to the isotonic regression used in the metric embedding (i.e., nonmetric MDS) step of the algorithm. We expect numerous applications to benefit from the presented embedding and visualization approach of PHATE, both in high throughput genomics and, more generally, in medical, empirical, and data sciences. Such additional applications will also be explored in future works.

5 Methods

5.1 Generation of Human EB Data

Low passage H1 hESCs were maintained on Matrigel-coated dishes in DMEM/F12-N2B27 media supplemented with FGF2. For EB formation, cells were treated with Dispase, dissociated into small clumps and plated in non-adherent plates in media supplemented with 20% FBS, which was prescreened for EB differentiation. Samples were collected during 3-day intervals during a 27 day-long differentiation timecourse. An undifferentiated hESC sample was also included (Fig. 16). Induction of key germ layer markers in these EB cultures was validated by qPCR (data not shown). For single cell analyses, EB cultures were dissociated, FACS sorted to remove doublets and dead cells and processed on a 10x genomics instrument to generate cDNA libraries, which were then sequenced. Small scale sequencing determined that we have successfully collected data on approximately 31,000 cells equally distributed throughout the timecourse.

5.2 Construction of the Artificial Tree Test Case

The artificial tree data shown in Fig. 3 is constructed using diffusion limited aggregation [11]. Branches are generated one at a time. A random point on the tree is chosen as the starting point of the new branch. The next branch is then generated. This process is repeated until the entire

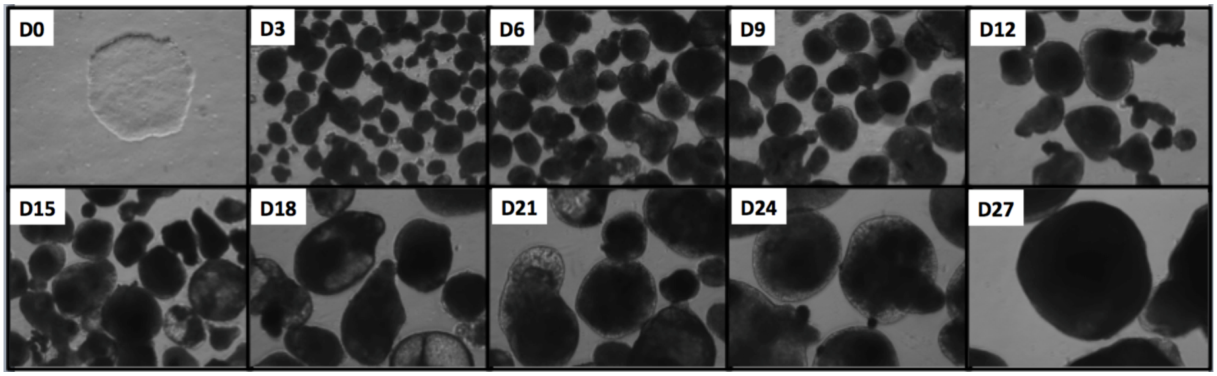

Figure 16: Inverted images of hESCs and EBs at each timepoint of data collection. Structures of different densities are clearly visible late in the time course (D15-D27) indicating the formation of distinct cell types.

tree is generated. We generate a tree with 20 distinct branches in 100 dimensions and 100 data points per branch. We then add zero mean Gaussian noise with standard deviation equal to 4.

5.3 Data Processing

In this section, we discuss methods we used to pre-process the data.

Data Subsampling The current PHATE implementation scales well for sample sizes up to approximately $N = 50000$. For N much larger than 50000, computational complexity can become an issue due to the multiple matrix operations required. All of the scRNAseq datasets considered in this paper have $N < 50000$. Thus, we used the full data and did not subsample these datasets. However, the mass cytometry datasets have much larger sample sizes. Thus, we randomly subsampled these datasets using uniform subsampling. The PHATE embedding is robust to the number of samples chosen, which we demonstrate later in the paper.

Mass Cytometry Data Preprocessing We process the mass cytometry datasets according to [9].

Single Cell RNA Sequencing Data Preprocessing This data was processed from raw reads to molecule counts using the Cell Ranger pipeline [69] Additionally, to minimize the effects of experimental artifacts on our analysis, we preprocess the scRNAseq data. We first perform library size normalization on the cells. scRNAseq data have large cell-to-cell variations in the number of observed molecules in each cell or *library size*. Some cells are highly sampled with many transcripts, while other cells are sampled with fewer. This variation is often caused by technical variations due to enzymatic steps including lysis efficiency, mRNA capture efficiency, and the efficiency of multiple amplification rounds [70]. Normalizing by the library size helps to correct for these technical variations. Normalization is accomplished by dividing the expression level of each gene in a cell by the library size of the corresponding cell.

After normalizing by the library size, we perform PCA to improve the robustness and reliability of the constructed affinity matrix $K_{k,\alpha}$. We choose the number of principal components to retain approximately 70% of the variance in the data which results in 20-50 principal components. We then take the log transform of the data.

Gut Microbiome Data Preprocessing We use the cleaned L6 American Gut data and remove samples that are near duplicates of other samples. We then preprocess the data using a similar approach for scRNAseq data. We first perform “library size” normalization to account for technical variations in different samples. We then use PCA to reduce the data to 30 dimensions and then log transform the data.

Applying PHATE to this data reveals several outlier samples that are very far from the rest of the data. We remove these samples and then reapply PHATE to the log-transformed data to obtain the results in Fig. 15.

5.4 Robustness Analysis of PHATE

In this section, we investigate the robustness of the PHATE embedding to subsampling and the choice of t.

Robustness to Subsampling We demonstrate that the PHATE algorithm is robust to subsampling of the data by running PHATE on the mass cytometry bone marrow dataset with varying subsample sizes N. The PHATE embedding for $N = 10000$ is shown in Fig. 3B while Fig. 17 shows the PHATE embedding for $N = 1000, 2500, 5000, 7500$. Note that most of the branches or trajectories that are visible when $N = 10000$ are still visible when $N = 7500, 5000$, and 2500. Even when $N = 1000$, several branches are still visible in the embedding. Thus, PHATE is robust to the subsampling size. Similar results can be obtained on other datasets.

Robustness to t In the Results section, we used the VNE to guide the choice of t in the PHATE embedding. Here, we show that the PHATE embedding is quite robust to the choice of t. Figure 18 shows the PHATE embedding on the bone marrow mass cytometry dataset with varying scale parameter t. Note that in Fig. 3B, we choose $t = 100$ for the embedding. Figure 18 shows that the embeddings for $85 \leq t \leq 115$ are nearly identical. Additionally, the embeddings for $t = 50$ and $t = 150$ are very similar to the embedding for $t = 100$. Thus, PHATE is also very robust to the scale parameter t. Similar results can be obtained on other datasets.

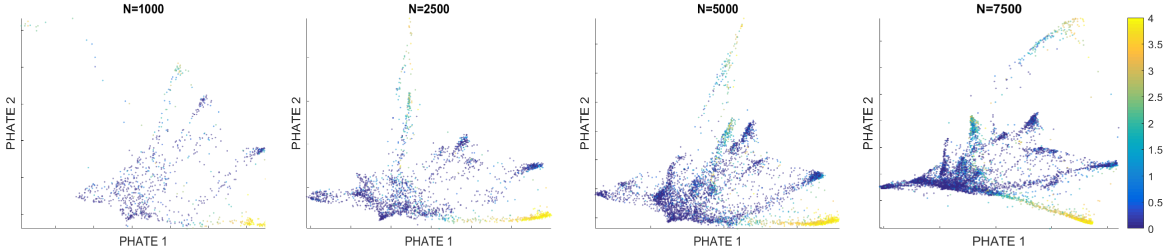

Figure 17: The PHATE embedding for the bone marrow mass cytometry dataset with varying number of subsample sizes N. The coloring corresponds to CD4 expression level. Most branches present for $N = 7500$ are also visible when $N = 5000$ or $N = 2500$ while several branches are still visible for even $N = 1000$, demonstrating that the PHATE embedding is robust to the size of the subsample. See also Fig. 3B for $N = 10000$.

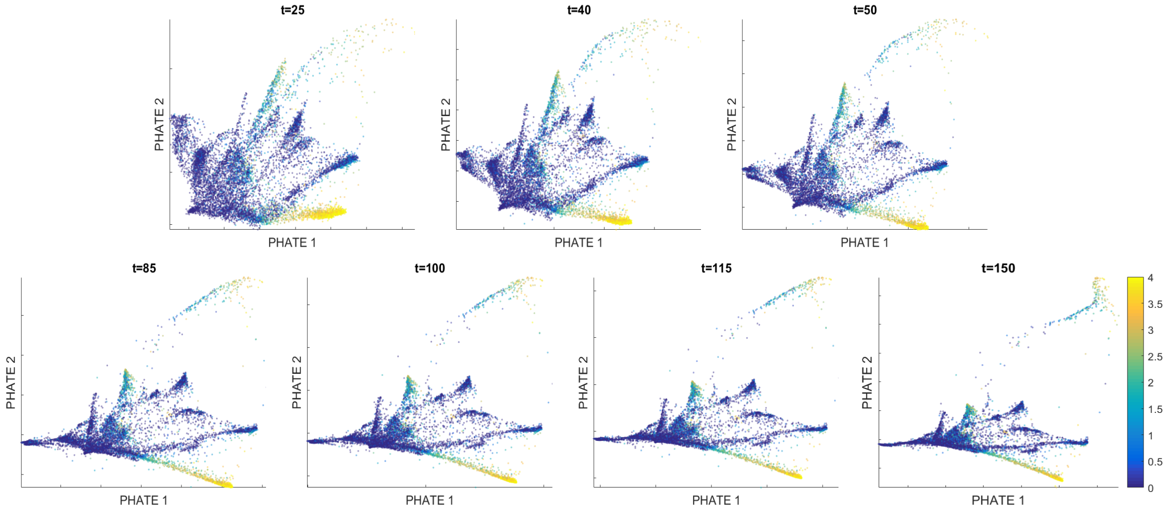

Figure 18: The PHATE embedding for the bone marrow mass cytometry dataset with varying scale parameter t. The embeddings for $85 \leq t \leq 115$ are nearly identical while the embeddings for $t = 50$ and $t = 150$ are still very similar to the embedding for $t = 100$. This demonstrates that the embedding is robust to the choice of t.

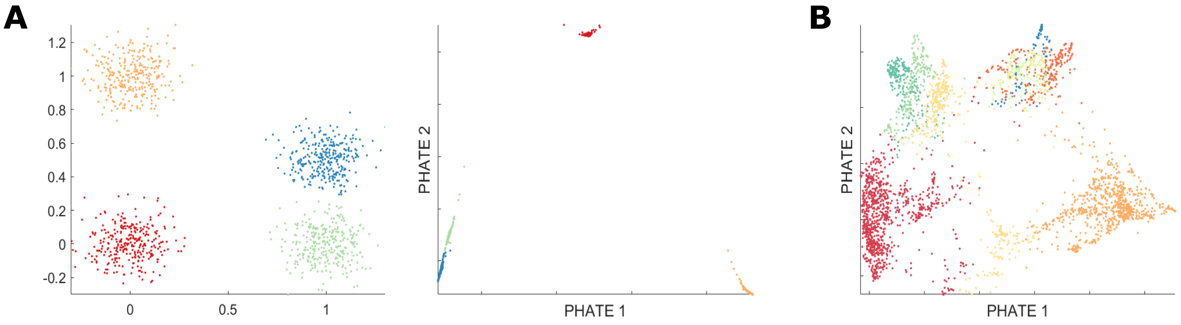

Figure 19: Effect of PHATE on naturally clustered data. (**A**) Left: samples from four 24-dimensional Gaussian distributions with identical covariance matrices ($\Sigma = 0.1I_{24}$ where I_d is a d-dimensional identity matrix) and different means: $\mu_1 = (0, 0, \dots, 0)^T$, $\mu_2 = (0, 1, 0, \dots, 0)^T$, $\mu_3 = (1, 0, \dots, 0)^T$, and $\mu_4 = (1, 0.5, 0, \dots, 0)^T$. Only the first two dimensions are shown. Right: the PHATE embedding applied to the data. Clusters that are clearly separated are not connected in the embedding while clusters that are very close are connected. (**B**) PHATE applied to data from [71]. The data points are colored by clusters from spectral clustering. Again, the main clusters are fairly separated from each other in the embedding.

5.5 PHATE on Clusters

We show that the PHATE embedding does not artificially connect clusters that are well separated from each other. Figure 19A shows PHATE applied to simulated data from four 24-dimensional Gaussian distributions with identical covariance matrices ($\Sigma = 0.1I_{24}$ where I_d is a d-dimensional identity matrix) and different means: $\mu_1 = (0, 0, \dots, 0)^T$, $\mu_2 = (0, 1, 0, \dots, 0)^T$, $\mu_3 = (1, 0, \dots, 0)^T$, and $\mu_4 = (1, 0.5, 0, \dots, 0)^T$. Two of the data clusters are linearly separable from each other and from the other two clusters, which have some overlap with each other. When PHATE is applied to the data, the separable clusters are still separable in the PHATE dimensions while the overlapping clusters are close to each other.

In Fig. 19B, we apply PHATE to data from [71] which has a natural clustering structure. Note that PHATE keeps the main clusters fairly separate from each other. This demonstrates that PHATE does not artificially connect clusters.

Software Software for PHATE are available via github for academic use:
https://github.com/SmitaKrishnaswamy/PHATE.

References

[1] L. v. d. Maaten and G. Hinton, "Visualizing data using t-sne," *Journal of Machine Learning Research*, vol. 9, no. Nov, pp. 2579–2605, 2008.

[2] X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H. Pliner, and C. Trapnell, "Reversed graph embedding resolves complex single-cell developmental trajectories," *bioRxiv*, p. 110668, 2017.

[3] P. Qiu, E. F. Simonds, S. C. Bendall, K. D. Gibbs Jr, R. V. Bruggner, M. D. Linderman, K. Sachs, G. P. Nolan, and S. K. Plevritis, "Extracting a cellular hierarchy from high-dimensional cytometry data with spade," *Nature biotechnology*, vol. 29, no. 10, pp. 886–891, 2011.

[4] S. C. Bendall, K. L. Davis, E.-a. D. Amir, M. D. Tadmor, E. F. Simonds, T. J. Chen, D. K. Shenfeld, G. P. Nolan, and D. Pe'er, "Single-cell trajectory detection uncovers progression and regulatory coordination in human b cell development," *Cell*, vol. 157, no. 3, pp. 714–725, 2014.

[5] M. Setty, M. D. Tadmor, S. Reich-Zeliger, O. Angel, T. M. Salame, P. Kathail, K. Choi, S. Bendall, N. Friedman, and D. Pe'er, "Wishbone identifies bifurcating developmental trajectories from single-cell data," *Nature biotechnology*, vol. 34, no. 6, pp. 637–645, 2016.

[6] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems*, pp. 3111–3119, 2013.

[7] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation.," in *The Conference on Empirical Methods in Natural Language Processing*, vol. 14, pp. 1532–1543, 2014.

[8] F. Paul, Y. Arkin, A. Giladi, D. A. Jaitin, E. Kenigsberg, H. Keren-Shaul, D. Winter, D. Lara-Astiaso, M. Gury, A. Weiner, *et al.*, "Transcriptional heterogeneity and lineage commitment in myeloid progenitors," *Cell*, vol. 163, no. 7, pp. 1663–1677, 2015.

[9] S. C. Bendall, E. F. Simonds, P. Qiu, D. A. El-ad, P. O. Krutzik, R. Finck, R. V. Bruggner, R. Melamed, A. Trejo, O. I. Ornatsky, *et al.*, "Single-cell mass cytometry of differential immune and drug responses across a human hematopoietic continuum," *Science*, vol. 332, no. 6030, pp. 687–696, 2011.

[10] E. R. Zunder, E. Lujan, Y. Goltsev, M. Wernig, and G. P. Nolan, "A continuous molecular roadmap to iPSC reprogramming through progression analysis of single-cell mass cytometry," *Cell Stem Cell*, vol. 16, no. 3, pp. 323–337, 2015.

[11] T. A. Witten and L. M. Sander, "Diffusion-limited aggregation," *Physical Review B*, vol. 27, no. 9, p. 5686, 1983.

[12] P. Bérard, G. Besson, and S. Gallot, "Embedding riemannian manifolds by their heat kernel," *Geometric and Functional Analysis*, vol. 4, no. 4, pp. 373–398, 1994.

[13] P. W. Jones, M. Maggioni, and R. Schul, "Manifold parametrizations by eigenfunctions of the laplacian and heat kernels," *Proceedings of the National Academy of Sciences*, vol. 105, no. 6, pp. 1803–1808, 2008.

[14] R. R. Coifman and S. Lafon, "Diffusion maps," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 5–30, 2006.

[15] B. Nadler, S. Lafon, R. Coifman, and I. Kevrekidis, "Diffusion maps, spectral clustering and eigenfunctions of fokker-planck operators," in *Advances in Neural Information Processing Systems*, pp. 955–962, 2005.

[16] B. Nadler, S. Lafon, R. R. Coifman, and I. G. Kevrekidis, "Diffusion maps, spectral clustering and reaction coordinates of dynamical systems," *Applied and Computational Harmonic Analysis*, vol. 21, no. 1, pp. 113–127, 2006.

[17] S. Butterworth, "On the theory of filter amplifiers," *Wireless Engineer*, vol. 7, no. 6, pp. 536–541, 1930.

[18] J. Neumann, *Mathematische grundlagen der quantenmechanik*. Verlag von Julius Springer Berlin, 1932.

[19] K. Anand, G. Bianconi, and S. Severini, "Shannon and von neumann entropy of random networks with heterogeneous expected degree," *Physical Review E*, vol. 83, no. 3, p. 036109, 2011.

[20] T. F. Cox and M. A. A. Cox, *Multidimensional Scaling*. Chapman & Hall/CRC, 2 ed., 2001.

[21] R. Coifman, Y. Shkolnisky, F. Sigworth, and A. Singer, "Graph laplacian tomography from unknown random projections," *IEEE Transactions on Image Processing*, vol. 17, no. 10, pp. 1891–1899, 2008.

[22] R. R. Lederman and R. Talmon, "Learning the geometry of common latent variables using alternating-diffusion," *Applied and Computational Harmonic Analysis*, 2015.

[23] J. B. Kruskal, "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis," *Psychometrika*, vol. 29, no. 1, pp. 1–27, 1964.

[24] J. B. Kruskal, "Nonmetric multidimensional scaling: A numerical method," *Psychometrika*, vol. 29, no. 2, pp. 115–129, 1964.

[25] J. B. Kruskal and M. Wish, *Multidimensional Scaling*, vol. 11. Sage, 1978.

[26] L. Haghverdi, M. Buettner, F. A. Wolf, F. Buettner, and F. J. Theis, "Diffusion pseudotime robustly reconstructs lineage branching," *Nature Methods*, vol. 13, no. 10, pp. 845–848, 2016.

[27] E.-a. D. Amir, K. L. Davis, M. D. Tadmor, E. F. Simonds, J. H. Levine, S. C. Bendall, D. K. Shenfeld, S. Krishnaswamy, G. P. Nolan, and D. Pe'er, "visne enables visualization of high dimensional single-cell data and reveals phenotypic heterogeneity of leukemia," *Nature biotechnology*, vol. 31, no. 6, pp. 545–552, 2013.

[28] K. M. Carter, R. Raich, and A. O. Hero III, "On local intrinsic dimension estimation and its applications," *IEEE Transactions on Signal Processing*, vol. 58, no. 2, pp. 650–663, 2010.

[29] K. R. Moon, J. J. Li, V. Delouille, R. De Visscher, F. Watson, and A. O. Hero, "Image patch analysis of sunspots and active regions-i. intrinsic dimension and correlation analysis," *Journal of Space Weather and Space Climate*, vol. 6, p. A2, 2016.

[30] J. A. Costa and A. O. Hero III, "Determining intrinsic dimension and entropy of high-dimensional shape spaces," in *Statistics and Analysis of Shapes*, pp. 231–252, Springer, 2006.

[31] P. Bonacich, "Factoring and weighting approaches to status scores and clique identification," *Journal of Mathematical Sociology*, vol. 2, no. 1, pp. 113–120, 1972.

[32] G. David and A. Averbuch, "Hierarchical data organization, clustering and denoising via localized diffusion folders," *Applied and Computational Harmonic Analysis*, vol. 33, no. 1, pp. 1–23, 2012.

[33] S. Krishnaswamy, M. H. Spitzer, M. Mingueneau, S. C. Bendall, O. Litvin, E. Stone, D. Pe'er, and G. P. Nolan, "Conditional density-based analysis of T cell signaling in single-cell data," *Science*, vol. 346, no. 6213, p. 1250689, 2014.

[34] D. van Dijk, J. Nainys, R. Sharma, P. Kathail, A. J. Carr, K. R. Moon, L. Mazutis, G. Wolf, S. Krishnaswamy, and D. Pe'er, "Magic: A diffusion-based imputation method reveals gene-gene interactions in single-cell rna-sequencing data," *bioRxiv*, p. 111591, 2017.

[35] H.-Y. Yang, D. K. Jeong, S.-H. Kim, K.-J. Chung, E.-J. Cho, C. H. Jin, U. Yang, S. R. Lee, D.-S. Lee, and T.-H. Lee, "Gene expression profiling related to the enhanced erythropoiesis in mouse bone marrow cells," *Journal of cellular biochemistry*, vol. 104, no. 1, pp. 295–303, 2008.

[36] J. D. Crispino, "Gata1 in normal and malignant hematopoiesis," in *Seminars in cell & developmental biology*, vol. 16, pp. 137–147, Elsevier, 2005.

[37] Y. Fujiwara, C. P. Browne, K. Cunniff, S. C. Goff, and S. H. Orkin, "Arrested development of embryonic red cell precursors in mouse embryos lacking transcription factor gata-1," *Proceedings of the National Academy of Sciences*, vol. 93, no. 22, pp. 12355–12358, 1996.

[38] L. Pevny, M. C. Simon, *et al.*, “Erythroid differentiation in chimaeric mice blocked by a targeted mutation in the gene for transcription factor gata-1,” *Nature*, vol. 349, no. 6306, p. 257, 1991.

[39] K. Fiolka, R. Hertzano, L. Vassen, H. Zeng, O. Hermesh, K. B. Avraham, U. Dührsen, and T. Möröy, “Gfi1 and gfi1b act equivalently in haematopoiesis, but have distinct, non-overlapping functions in inner ear development,” *EMBO reports*, vol. 7, no. 3, pp. 326–333, 2006.

[40] L. Van der Meer, J. Jansen, and B. Van Der Reijden, “Gfi1 and gfi1b: key regulators of hematopoiesis,” *Leukemia*, vol. 24, no. 11, pp. 1834–1843, 2010.

[41] H.-Y. Yang, S. H. Kim, S.-H. Kim, D.-J. Kim, S.-U. Kim, D.-Y. Yu, Y. I. Yeom, D.-S. Lee, Y.-J. Kim, B.-J. Park, *et al.*, “The suppression of zfpm-1 accelerates the erythropoietic differentiation of human cd34+ cells,” *Biochemical and biophysical research communications*, vol. 353, no. 4, pp. 978–984, 2007.

[42] J. M. Polo, E. Anderssen, R. M. Walsh, B. A. Schwarz, C. M. Nefzger, S. M. Lim, M. Borkent, E. Apostolou, S. Alaei, J. Cloutier, *et al.*, “A molecular roadmap of reprogramming somatic cells into ips cells,” *Cell*, vol. 151, no. 7, pp. 1617–1632, 2012.

[43] H.-P. Huang, P.-H. Chen, C.-Y. Yu, C.-Y. Chuang, L. Stone, W.-C. Hsiao, C.-L. Li, S.-C. Tsai, K.-Y. Chen, H.-F. Chen, *et al.*, “Epithelial cell adhesion molecule (epcam) complex proteins promote transcription factor-mediated pluripotency reprogramming,” *Journal of Biological Chemistry*, vol. 286, no. 38, pp. 33520–33532, 2011.

[44] K. Takahashi and S. Yamanaka, “Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors,” *cell*, vol. 126, no. 4, pp. 663–676, 2006.

[45] H. Hong, K. Takahashi, T. Ichisaka, T. Aoi, O. Kanagawa, M. Nakagawa, K. Okita, and S. Yamanaka, “Suppression of induced pluripotent stem cell generation by the p53–p21 pathway,” *Nature*, vol. 460, no. 7259, pp. 1132–1135, 2009.

[46] M. Bibel, J. Richter, E. Lacroix, and Y.-A. Barde, “Generation of a defined and uniform population of cns progenitors and neurons from mouse embryonic stem cells,” *Nature protocols*, vol. 2, no. 5, pp. 1034–1043, 2007.

[47] S.-M. Kang, M. S. Cho, H. Seo, C. J. Yoon, S. K. Oh, Y. M. Choi, and D.-W. Kim, “Efficient induction of oligodendrocytes from human embryonic stem cells,” *Stem Cells*, vol. 25, no. 2, pp. 419–424, 2007.

[48] X. Zhao, J. Liu, and I. Ahmad, “Differentiation of embryonic stem cells to retinal cells in vitro,” *Embryonic Stem Cell Protocols: Volume 2: Differentiation Models*, pp. 401–416, 2006.

[49] S. S. Liour, S. A. Kraemer, M. B. Dinkins, C.-Y. Su, M. Yanagisawa, and R. K. Yu, "Further characterization of embryonic stem cell-derived radial glial cells," *Glia*, vol. 53, no. 1, pp. 43–56, 2006.

[50] T. Nakano, H. Kodama, and T. Honjo, "In vitro development of primitive and definitive erythrocytes from different precursors," *Science*, vol. 272, no. 5262, p. 722, 1996.

[51] S.-I. Nishikawa, S. Nishikawa, M. Hirashima, N. Matsuyoshi, and H. Kodama, "Progressive lineage analysis by cell sorting and culture identifies flk1+ ve-cadherin+ cells at a diverging point of endothelial and hemopoietic lineages," *Development*, vol. 125, no. 9, pp. 1747–1757, 1998.

[52] M. V. Wiles and G. Keller, "Multiple hematopoietic lineages develop from embryonic stem (es) cells in culture," *Development*, vol. 111, no. 2, pp. 259–267, 1991.

[53] A. J. Potocnik, P. J. Nielsen, and K. Eichmann, "In vitro generation of lymphoid precursors from embryonic stem cells.," *The EMBO journal*, vol. 13, no. 22, p. 5274, 1994.

[54] M. Tsai, J. Wedemeyer, S. Ganiatsas, S.-Y. Tam, L. I. Zon, and S. J. Galli, "In vivo immunological function of mast cells derived from embryonic stem cells: an approach for the rapid analysis of even embryonic lethal mutations in adult mice in vivo," *Proceedings of the National Academy of Sciences*, vol. 97, no. 16, pp. 9186–9190, 2000.

[55] P. Fairchild, F. Brook, R. Gardner, L. Graca, V. Strong, Y. Tone, M. Tone, K. Nolan, and H. Waldmann, "Directed differentiation of dendritic cells from mouse embryonic stem cells," *Current Biology*, vol. 10, no. 23, pp. 1515–1518, 2000.

[56] J. Yamashita, H. Itoh, M. Hirashima, M. Ogawa, S. Nishikawa, T. Yurugi, M. Naito, K. Nakao, and S.-I. Nishikawa, "Flk1-positive cells derived from embryonic stem cells serve as vascular progenitors," *Nature*, vol. 408, no. 6808, pp. 92–96, 2000.

[57] V. A. Maltsev, J. Rohwedel, J. Hescheler, and A. M. Wobus, "Embryonic stem cells differentiate in vitro into cardiomyocytes representing sinusnodal, atrial and ventricular cell types," *Mechanisms of development*, vol. 44, no. 1, pp. 41–50, 1993.

[58] J. Rohwedel, V. Maltsev, E. Bober, H.-H. Arnold, J. Hescheler, and A. Wobus, "Muscle cell differentiation of embryonic stem cells reflects myogenesis in vivo: developmentally regulated expression of myogenic determination genes and functional expression of ionic currents," *Developmental biology*, vol. 164, no. 1, pp. 87–101, 1994.

[59] G. Kania, P. Blyszczuk, A. Jochheim, M. Ott, and A. M. Wobus, "Generation of glycogen- and albumin-producing hepatocyte-like cells from embryonic stem cells," *Biological chemistry*, vol. 385, no. 10, pp. 943–953, 2004.

[60] I. S. Schroeder, A. Rolletschek, P. Blyszczuk, G. Kania, and A. M. Wobus, "Differentiation of mouse embryonic stem cells to insulin-producing cells," *Nature Protocols*, vol. 1, no. 2, pp. 495–507, 2006.

[61] N. Geijsen, M. Horoschak, K. Kim, J. Gribnau, K. Eggan, and G. Q. Daley, "Derivation of embryonic germ cells and male gametes from embryonic stem cells," *Nature*, vol. 427, no. 6970, pp. 148–154, 2004.

[62] J. Kehler, K. Hübner, S. Garrett, and H. R. Schöler, "Generating oocytes and sperm from embryonic stem cells," *Seminars in reproductive medicine*, vol. 23, no. 03, pp. 222–233, 2005.

[63] N. Patterson, P. Moorjani, Y. Luo, S. Mallick, N. Rohland, Y. Zhan, T. Genschoreck, T. Webster, and D. Reich, "Ancient admixture in human history," *Genetics*, vol. 192, no. 3, pp. 1065–1093, 2012.

[64] N. Patterson, A. L. Price, and D. Reich, "Population structure and eigenanalysis," *PLoS genet*, vol. 2, no. 12, p. e190, 2006.

[65] M. Arumugam, J. Raes, E. Pelletier, D. Le Paslier, T. Yamada, D. R. Mende, G. R. Fernandes, J. Tap, T. Bruls, J.-M. Batto, M. Bertalan, N. Borruel, F. Casellas, L. Fernandez, L. Gautier, T. Hansen, M. Hattori, T. Hayashi, M. Kleerebezem, K. Kurokawa, M. Leclerc, F. Levenez, C. Manichanh, H. Nielsen, T. Nielsen, N. Pons, J. Poulain, J. Qin, T. Sicheritz-Ponten, S. Tims, D. Torrents, E. Ugarte, E. Zoetendal, J. Wang, F. Guarner, O. Pedersen, W. de Vos, S. Brunak, J. Dore, MetaHIT Consortium, J. Weissenbach, S. Ehrlich, and P. Bork, "Enterotypes of the human gut microbiome," *Nature*, vol. 473, no. 7346, pp. 174–180, 2011.

[66] D. McDonald, A. Birmingham, and R. Knight, "Context and the human microbiome," *Microbiome*, vol. 3, no. 1, p. 52, 2015.

[67] J. D. Silverman, A. Washburne, S. Mukherjee, and L. A. David, "A phylogenetic transform enhances analysis of compositional microbiota data," *eLife*, 2017.

[68] G. D. Wu, J. Chen, C. Hoffmann, K. Bittinger, Y.-Y. Chen, S. A. Keilbaugh, M. Bewtra, D. Knights, W. A. Walters, R. Knight, *et al.*, "Linking long-term dietary patterns with gut microbial enterotypes," *Science*, vol. 334, no. 6052, pp. 105–108, 2011.

[69] G. X. Zheng, J. M. Terry, P. Belgrader, P. Ryvkin, Z. W. Bent, R. Wilson, S. B. Ziraldo, T. D. Wheeler, G. P. McDermott, J. Zhu, M. Gregory, J. Shuga, L. Montesclaros, J. Underwood, D. Masquelier, S. Nishimura, M. Schnall-Levin, P. Wyatt, C. Hindson, R. Bharadwaj, A. Wong, K. D. Ness, L. W. Beppu, H. J. Deeg, C. McFarland, K. R. Loeb, W. J. Valente, N. G. Ericson, E. A. Stevens, J. P. Radich, T. S. Mikkelsen, B. J. Hindson, and J. H. Bielas, "Massively parallel digital transcriptional profiling of single cells," *Nature Communications*, vol. 8, p. 14049, 2017.

[70] D. Grün, L. Kester, and A. Van Oudenaarden, “Validation of noise models for single-cell transcriptomics,” *Nature methods*, vol. 11, no. 6, pp. 637–640, 2014.

[71] A. Zeisel, A. B. Muñoz-Manchado, S. Codeluppi, P. Lönnerberg, G. La Manno, A. Juréus, S. Marques, H. Munguba, L. He, C. Betsholtz, *et al.*, “Cell types in the mouse cortex and hippocampus revealed by single-cell rna-seq,” *Science*, vol. 347, no. 6226, pp. 1138–1142, 2015.