Visualizing Structure and Transitions for Biological Data Exploration

Kevin R. Moon[1,2,3†], David van Dijk[1,3†], Zheng Wang[4†], Scott Gigante[1], Daniel Burkhardt[1], William S. Chen[1], Kristina Yim[1], Antonia van den Elzen[1], Matthew J. Hirn[5,6], Ronald R. Coifman[2], Natalia B. Ivanova[4‡**], Guy Wolf[2‡], Smita Krishnaswamy[1,3‡*]

[1]Department 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]Department of Computational Mathematics, Science and Engineering;
[6]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

With the advent of high-throughput technologies measuring high-dimensional biological data, there is a pressing need for visualization tools that reveal the structure and emergent patterns of data in an intuitive form. We present PHATE, a visualization method that captures both local and global nonlinear structure in data by an information-geometry distance between datapoints. We perform extensive comparison between PHATE and other tools on a variety of artificial and biological datasets, and find that it consistently preserves a range of patterns in data including continual progressions, branches, and clusters. We show that PHATE is applicable to a wide variety of datatypes including mass cytometry, single-cell RNA-sequencing, Hi-C, and gut microbiome data, where it can generate interpretable insights into the underlying systems. Finally, we use PHATE to explore a newly

generated scRNA-seq dataset of human germ layer differentiation. Here, PHATE reveals a dynamic picture of the main developmental branches in unparalleled detail.

1 Introduction

High dimensional, high-throughput data is accumulating at a staggering rate, especially from biological datasets featuring single-cell transcriptomics and other genomic and epigenetic assays. Since humans are visual learners, it is vitally important that these datasets are presented to researchers in intuitive ways, with faithful visualizations that can lead to hypothesis generation. However, few methods can produce clean and denoised visualizations that reveal both fine-grained local structure (relationships between near-neighbors) and global structure (large scale relationships between clusters and progressions) without imposing stringent assumptions on the underlying system on which the data is measured. This is especially important in biological systems where structure exists at many different scales and is often unknown.

Here, we present **PHATE (Potential of Heat-diffusion for Affinity-based Transition Embedding)**, a new type of method for visualizing high-dimensional data in a scalable manner (Figure 1). PHATE models the data as a statistical manifold where each datapoint is represents a probability distribution (constructed via data diffusion). PHATE then preserves informational distances that capture the intrinsic geometry of the dataset. The result is that high-dimensional and non-linear structures such as clusters, non-linear progressions, and branches become apparent in two or three dimensions and can be extracted for further analysis (Figure 1A).

There are currently two types of methods that are used for visually distilling structure from high-throughput data. These include:

1. Model fitting methods where specific pre-determined structures such as clusters, graphs, or hierarchical trees are fitted to approximate the data. Such methods rely on assumptions made by biologists regarding the shape and structure of the data. Examples of such methods include Wanderlust (which assumes a single-trajectory structure) [1], Wishbone (which assumes a single Y branch) [2], Diffusion Pseudotime (which assumes a branching tree) [3], and Monocle 2 (which assumes a tree) [4].

2. Distance or neighborhood-preserving embeddings that aim to preserve some notion of similarity or distance in lower dimensional coordinates. For instance, PCA attempts to retain global variance while t-SNE attempts to retain "stochastic neighborhood structure" [5].

PHATE belongs to the second class of methods and does not make any assumptions on the structure of the data. Thus PHATE is especially useful for datasets which measure new types of systems where the underlying structure is unknown.

However, existing distance or neighborhood preserving methods such as t-SNE, PCA, Isomap [6], and Diffusion Maps [7] are limited in their ability to visualize and preserve structures of interest in biology. For instance, t-SNE only preserves local neighbors leading to embeddings that

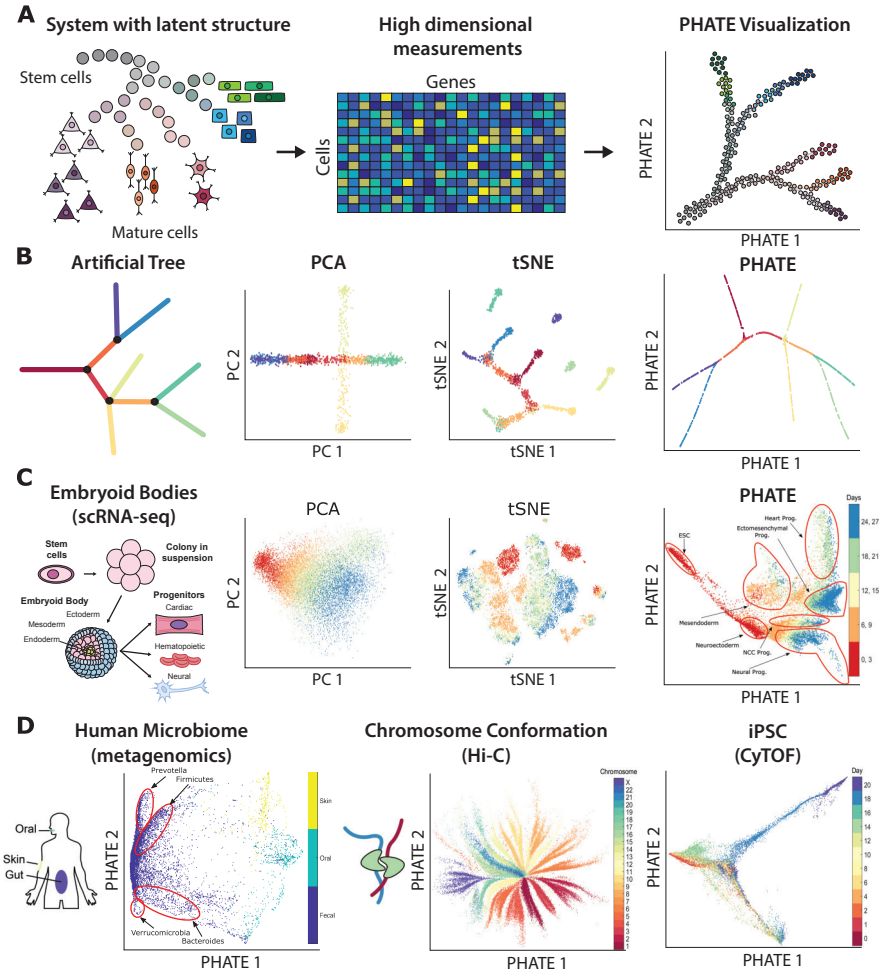


Figure 1: Overview of PHATE and its ability to reveal structure in data. (**A**) Conceptual figure demonstrating the progression of stem cells into different cell types and the corresponding high dimensional single-cell measurements rendered as a visualization by PHATE. (**B**) (Left) A 2D drawing of an artificial tree with color-coded branches. Data is uniformly sampled from each branch in 60 dimensions with Gaussian noise added (see Methods). (Right) Comparison of PCA, t-SNE, and the PHATE visualizations for the high-dimensional artificial tree data. PHATE is best at revealing global and branching structure in the data. In particular, PCA cannot reveal fine-grained local features such as branches while t-SNE breaks the structure apart and shuffles the broken pieces within the visualization. See Figures 4 and S1 for more comparisons on artificial data. (**C**) Comparison of PCA, t-SNE, and the PHATE visualizations for new embryoid body data showing similar trends as in (B). (**D**) PHATE applied to various datatypes. Left: PHATE on human microbiome data shows clear distinctions between skin, oral and fecal samples, as well as different enterotypes within the fecal samples. Middle: PHATE on Hi-C chromatin conformation data shows the global structure of chromatin. The embedding is colored by the different chromosomes. Right: PHATE on induced pluripotent stem cell (iPSC) CyTOF data. The embedding is colored by time after induction. See Figures 5, 6, and S1 for more applications to real data.

"shatter" the space (Figure 1B-C). Moreover, t-SNE has a KL-divergence penalty that effectively penalizes the placement of points relative to close neighbors and not relative to far-away points, i.e., it misses global structure. Thus, even when neighborhoods are extended via diffusion or other means, the t-SNE does not penalize placing far objects near each other. On the other hand, PCA is a linear method that only preserves global variance and cannot denoise in nonlinear directions, leading to noisy embeddings (Figure 1B-C). Diffusion Maps can follow and denoise in non-linear directions but they are not amenable to low-dimensional embeddings for visualization as we explain below.

By contrast, PHATE defines a new notion of distance we term *potential distance* that, when preserved in low dimensions via metric MDS, solves many of these issues, i.e., produces a denoised embedding that captures both local and global manifold-intrinsic distances to reveal structure. To develop this potential distance, we used the combination of diffusion-based manifold learning, stochastic dynamics, as well as information theory: First, a diffusion operator is used to re-represent each datapoint (e.g. cells in single-cell data) as a probability distribution of walking probabilities. Then, a pairwise distance is derived by taking the potential of the data diffusion process, and using a novel information geometric distance (which is also a divergence) between potentials.

The connection between diffusion geometry and information geometry has been explored in previous work [8, 9] which show that diffusion maps can be used to empirically estimate an information geometry of dynamical systems from noisy samples. Here, we augment and complete this relation by establishing that, conversely, information geometry tools can be used to analyze and embed diffusion geometry as an alternative to classic diffusion maps—an alternative that is better-suited for low dimensional visualization.

We show that PHATE consistently outperforms existing methods on a wide variety of benchmark test cases where the ground truth is known. We also use PHATE to visualize several biological and non-biological real world datasets (Figure 1D). To demonstrate the ability of PHATE to reveal new biological insights, we apply PHATE to a newly generated single-cell RNA-sequencing dataset of human embryonic stem cells grown as embryoid bodies over a period of 27 days to observe differentiation into diverse cell lineages. PHATE successfully captures all branches of development within this system and enables the isolation of rare populations based on surface markers, which we experimentally validate.

2 The PHATE Algorithm

PHATE uses a series of carefully-crafted steps to: 1. Re-represent a cell as a probability distribution of affinities, 2. derive a distance between these distributions with desirable properties, 3. Embed the data optimally in low dimensions for visualization, and 4. Automatically extract information about branchpoints from the visualization.

PHATE learns manifold-intrinsic affinities between datapoints via data diffusion. Diffusing through data is a concept that was popularized in the derivation of Diffusion Maps (DM) [7].

However, the concept of diffusion is older and harks back to random walks on graphs. To diffuse through data, PHATE transforms input data measurements (Figure 2A) to distances (Figure 2B), and then affinities (Figure 2C) using a carefully designed and rapidly decaying kernel. Then PHATE creates a diffusion operator by row normalizing the affinities. This diffusion operator is a row-stochastic matrix that contains the probabilities of transitioning from one data point to another in each single step of a Markovian random walk. Then, to reveal long-range nonlinear pathways in the data, PHATE propagates the local single-step transitions by powering the diffusion operator (Figure 2D) to a power t that maximizes signal dimensions and minimizes noise dimensions as indicated by a novel use of von Neumann entropy, which measures the information content of an operator. The powered operator integrates information from multiple random walks over the data. Therefore, it captures longer distance connections while also locally denoising the data. Such walks also have a smoothing (denoising) effect on the affinities, and amount to a low-pass filtering of the data graph. The use of a Markovian diffusion operator naturally results in each row (corresponding to a data point) being a probability distribution, thus creating a statistical manifold.

Next, PHATE derives a distance between two points x and y on this statistical manifold. We call this distance the *potential distance* (Figure 2E). It is calculated by computing the distance between log-transformed transition probabilities from the powered diffusion operator. The key insight in formulating the potential distance is that an informational distance between probability distributions is more sensitive to global relationships (between far-away points) and more stable at boundaries of manifolds than straight point-wise comparisons of probabilities (i.e., diffusion distances). This is because the diffusion distance is sensitive to differences between the main modes of the diffused probabilities and is largely insensitive to differences in the tails. In contrast, the potential distance, or more generally informational distances, use a submodular function (such as a log) to render distances sensitive to differences in both the main modes and the tails. This gives PHATE the ability to preserve both local and manifold-intrinsic global distances in a way that is optimized for visualization. The resulting metric space also quantifies differences between energy potentials that dominate "heat" propagation along diffusion pathways (i.e., based on the heat-equation diffusion model) between data points, instead of simply considering transition probabilities along them.

The potential distances are then preserved in low dimensions via metric MDS, which aims to create a low-dimensional representation such that the distances between points in the low-dimensional space approximate the corresponding potential distances. The goal is to preserve the maximal information about distances in a limited number of dimensions. Thus we use a variance-preserving embedding instead of taking the major directions of nonlinear variance in a diffusion operator. This allows PHATE to provide a low dimensional embedding for visualization in contrast with Diffusion Maps, which eigendecomposes the diffusion operator.

This is the full PHATE method, which is described in Algorithm 1. Computational aspects of each step are expanded in Methods.

In addition to the exact computation of PHATE, we have a fast version of PHATE that produces near-identical results. In this version, PHATE is implemented in an efficient and scalable

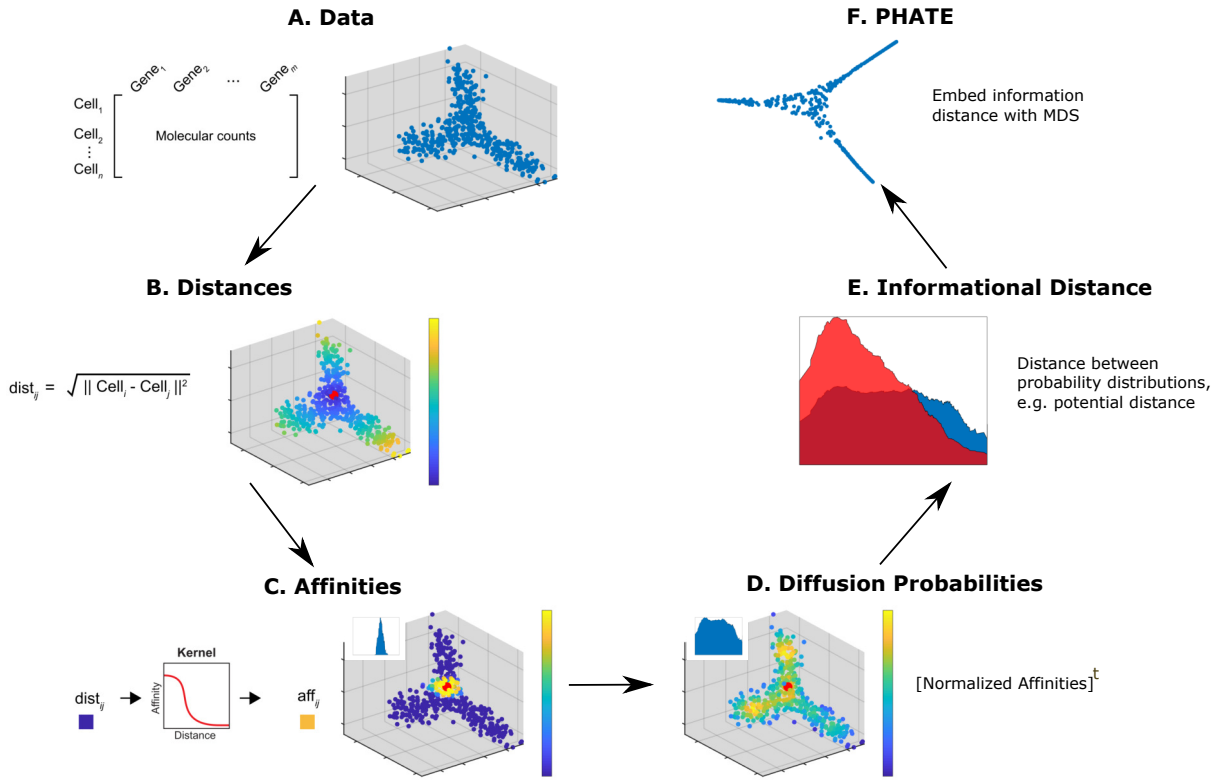


Figure 2: Steps of the PHATE algorithm. (**A**) Data. (**B**) Euclidean distances. Data points are colored by their Euclidean distance to the highlighted point. (**C**) Markov-normalized affinity matrix. Distances are transformed to local affinities via a kernel function and then normalized to a probability distribution. Data points are colored by the probability of transitioning from the highlighted point in a single step random walk. (**D**) Diffusion probabilities. The normalized affinities are diffused to denoise the data and learn long-range relationships between points. Data points are colored by the probability of transitioning from the highlighted point in a t step random walk. (**E**) Informational distance. An informational distance (e.g. the potential distance) that measures the dissimilarity between the diffused probabilities is computed. (**F**) The final PHATE embedding. The informational distances are embedded into low dimensions using MDS. Note that distances or affinities can be directly input to the appropriate step in cases of connectivity data. Therefore, the Euclidean distance or our constructed affinities can be replaced with distances or affinities that best describe the data. For example, in Figure 6F we replace our affinity matrix with the Facebook connectivity matrix.

manner by using landmark subsampling, sparse matrices, and randomized matrix decompositions. For more details on the scalability of PHATE see Section 6.1.8 in Methods, Algorithm 2, and Figure S5, which shows the fast runtime of PHATE on datasets of different sizes, including a dataset of 1.3 million cells (2.5 hours) and a network of 1.8 million nodes (12 minutes).

Algorithm 1: The PHATE algorithm

Input: Data matrix X, neighborhood size k, locality scale α, desired embedding dimension m (usually 2 or 3 for visualization)

Output: The PHATE embedding Y_m

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 ε_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: $\mathfrak{V}^t \leftarrow$ compute potential distance matrix from U_t (see Def 1)
9: $Y_{class} \leftarrow$ apply classical MDS to $\mathfrak{V}^t$
10: $Y_m \leftarrow$ apply metric MDS to $\mathfrak{V}^t$ with Y_{class} as an initialization

2.1 Steps of PHATE

The embedding provided by PHATE is specifically designed to enable visualization of global and local structure in exploratory settings with the following criteria in mind:

Visualization: To enable visualization, PHATE captures variance in low (2-3) dimensions.

Manifold-structure preserving: To provide an interpretable view of dynamics (e.g., pathways or progressions) in the data, PHATE preserves and emphasizes global nonlinear transitions in the data, in addition to local transitions.

Denoising: To enable unsupervised data exploration, PHATE denoises data such that progressions within the data are immediately identifiable and clearly separated.

Robust: PHATE produces a robust embedding in the sense that the revealed boundaries and the intersections of progressions within the data are insensitive to user configurations of the algorithm.

Here we explain how each of the steps in PHATE helps us ensure that the provided embedding satisfies the four properties described above.

Distances Consider the common approach of linearly embedding the raw data matrix itself, e.g., with PCA, to preserve the global structure of the data. PCA finds the directions of the data that capture the largest global variance. However, in most cases local transitions are noisy and global transitions are nonlinear. Therefore, linear notions such as global variance maximization are insufficient to capture latent patterns in the data, and they typically result in noisy visualization (Figure 4, Column 2). To provide reliable *structure preservation* that emphasizes transitions in the data, we need to consider the *intrinsic* structure of the data. This implies and motivates preserving distances between data points (e.g., cells) that consider gradual changes between them along these nonlinear transitions.

Affinities A standard choice of a distance metric is the Euclidean distance. However, global Euclidean distances are not reflective of transitions in the data, especially in biological datasets that have nonlinear and noisy structures. For instance, cells sampled from a developmental system, such as hematopoiesis or embryonic stem cell differentiation, show gradual changes where adjacent cells are only slightly different from each other. But these changes quickly aggregate into nonlinear transitions in marker expression along each developmental path. Therefore, we transform the global Euclidean distances into local affinities that quantify the similarities between nearby (in the Euclidean space) data points. We do this via a novel, sharply decreasing exponential kernel we designed, which we call the α-decaying kernel (Figure S2D), in conjunction with a locally adaptive bandwidth to handle data density variations (see Methods for details).

Propagating Affinities via Diffusion Embedding local affinities directly can result in a loss of global structure as is evident in t-SNE (Figures 1, 4, 5, and S1) or kernel PCA embeddings. For example, t-SNE only preserves data clusters, but not transitions between clusters, since it does not enforce any preservation of global structure. In contrast, a faithful structure-preserving embedding (and visualization) needs to go beyond local affinities (or distances), and consider more global relations between parts of the data. Therefore, to process the local affinities into global relations, we use the affinities to define small local steps between data points, and then chain them together to "walk" through the data. This process of propagating local affinities to global random walks is formulated by powering the Markov normalized affinity matrix, or the diffusion operator. This provides a global and robust intrinsic data distance that preserving the overall structure of the data.

In addition to learning the global structure, powering the diffusion operator has the effect of low-pass filtering the data such that the main pathways in it are emphasized and small noise dimensions are diminished, thus achieving the denoising objective of our method as well. In Figures 1B and 4A-B we see that the tree dataset successfully denoised by PHATE. We use this denoising effect to automatically select the diffusion time scale t. Intuitively, t is set to be the number of diffusion time-steps required to maximize the denoising aspect of PHATE while minimizing the loss of local structure information in the diffusion geometry. To assess this time-step, we propose a novel knee-point analysis of the spectral entropy (also known as Von

Neumann Entropy) of the powered diffusion operator at various time scales (Figure S2E). We note that the use of VNE effectively eliminates the diffusion parameter t, which may be useful for a variety of diffusion-based methods in addition to PHATE.

Embedding Potential Distances To resolve instabilities in diffusion distances and embed the global structure captured by the diffusion geometry in low (2 or 3) dimensions, we instead use a novel diffusion-based informational distance, which we call potential distance. The potential distance is inspired by information theory and stochastic dynamics, both fields where probability distributions are compared for different purposes. First, in information theory literature, divergences are utilized to measure discrepancies between probability distributions in the information space rather than the probability space, as they are more sensitive to differences between the tails of the distributions. Second, when analyzing dynamical systems of moving particles, it is not the point-wise difference between absolute particle counts that is used to compare states, but rather the ratio between these counts. Indeed, in the latter case the *Boltzmann Distribution Law* directly relates these ratios to differences in the energy of a state in the system. Therefore, similar to the information theory case, dynamical states are differentiated in energy terms, rather than probability terms. We employ the same reasoning in our case by defining our potential distance using localized diffusion energy potentials, rather than diffusion transition probabilities.

To go from the probability space to the energy (or information) space, we log transform the probabilities in the powered diffusion operator and consider an L^2 distance between these localized energy potentials in the data as our intrinsic data distance, which forms an M-divergence between the diffusion probability distributions [10, 11].

Alternative Informational Distances We note that the potential distance is a particular case of a wider family of diffusion-based informational distances that view the diffusion geometry as a statistical manifold in information geometry. We offer a generalization of these to a family of distances, where the diffusion distance is at one extreme of this family and the potential distance is at the other. Diffusion distances directly compare probability distributions pointwise with no *damping* of high probabilities. Thus changes in the tails of distributions do not contribute much to the distance. The potential distance is damped using the $\log$ function to the point where fold changes, even in low probabilities, have impact on the distance. Other distances, including the Hellinger distance [12], are in between these two. We introduce a parameter γ as a knob to control the level of damping, with the diffusion distance having $\gamma = -1$, the potential distance with $\gamma = 1$ and the Hellinger distance at $\gamma = 0$. See Section 6.1.5 in Methods for more details.

Figure S5H shows PHATE visualizations of the retinal bipolar data from [13] using different values of γ. This figure also indicates the impact of γ on the global structure captured by these distances, where indeed the global structure of the potential distance ($\gamma = 1$) is more similar (as compared to other γ values) to the structure captured by PCA, which is known to preserve global structure. Another way to see this is that the structure is unraveled less than when using diffusion distances.

Embedding in Low Dimensions A popular approach for embedding diffusion geometries is to use the eigendecomposition of the diffusion operator to build a *diffusion map* of the data. However, this approach tends to isolate progression trajectories into numerous diffusion coordinates (i.e., eigenvectors of the diffusion operator; see Figure S2B). In fact, this specific property was used in [3] as a heuristic for ordering cells along specific developmental tracks. Therefore, while diffusion maps preserve global structure and denoise the data, their higher intrinsic dimensionality is not amenable for visualization. Instead, we squeeze the variability into low dimensions using metric MDS (see Figure 2F).

A naïve approach towards obtaining a truly low dimensional embedding of diffusion geometries is to directly apply metric MDS, from the diffusion map space to a two dimensional space. However, as seen in Figure S1 (Column 5), direct embedding of this distance produces distorted visualizations. Embedding the potential distances is more stable at boundary conditions near end points compared to diffusion maps, even in the case of simple curves that contain no branching points. Figure S2C shows a half circle embedding with diffusion distances versus distances between log-scaled diffusion. We see that points are compressed towards the boundaries of the figure in the former. Additionally, this figure demonstrates that in the case of a full circle (i.e., with no end points or boundary conditions), our potential embedding (PHATE) yields the same representation as diffusion maps.

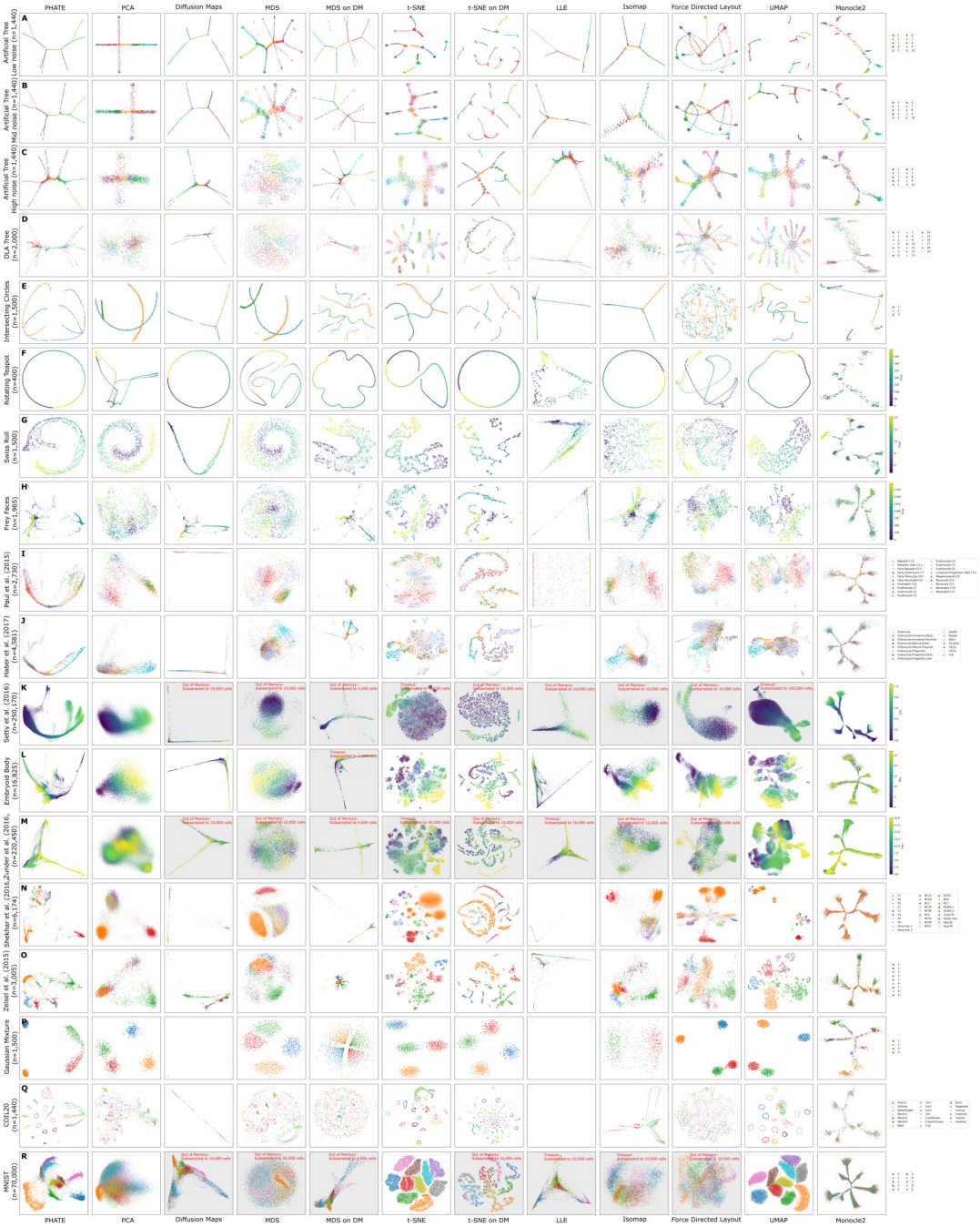
PHATE achieves an embedding that satisfies all four properties delineated above: PHATE preserves and emphasizes the global and local structure of the data via: 1. a localized affinity that is chained via diffusion to form global affinities through the intrinsic geometry of the data, 2. denoises the data by low-pass filtering through diffusion, 3. provides a distance that accounts for local and global relationships in the data and has robust boundary conditions for purposes of visualization, and 4. captures the data in low dimensions, using MDS, for visualization.

We have shown by demonstration in Figure S1 that all of the steps of PHATE, including the potential transform and MDS, are necessary, as diffusion maps, tSNE on diffusion maps, and MDS on diffusion maps fail to provide an adequate visualization in several benchmark test cases with known ground truth (even when using the same customized α-decaying kernel we developed for PHATE).

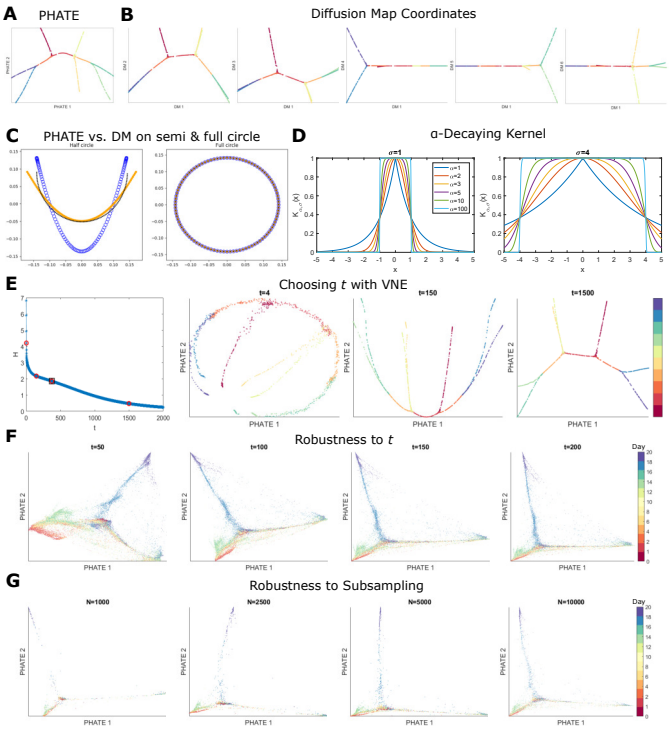
2.2 Extracting Information from PHATE

PHATE embeddings contain a large amount of information on the structure of the data, namely, local transitions, progressions, branches or splits in progressions, and end states of progression. In this section, we present new methods that provide suggested end points, branch points, and branches based on the information from higher dimensional PHATE embeddings. These may not always correspond to real decision points, but provide an annotation to aid the user in interpreting the PHATE visual.

Branch Point Identification with Local Intrinsic Dimensionality Since PHATE emphasizes progressions, PHATE plots can show branch points or divergences in progression. In



Supplemental Figure S1: Comparison of PHATE to various methods on multiple datasets. See Section 6.2 for a discussion.



Supplemental Figure S2: Impact of potential distances and PHATE parameters on the resulting visualization. (**A**) PHATE applied to the artificial tree data. Only two PHATE coordinates are needed to separate all branches. (**B**) The first six diffusion map coordinates of the artificial tree data. At least five of these coordinates are necessary to separate all of the branches. (**C**) 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. (**D**) 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 $\sigma = 1$ (left) and $\sigma = 4$ (right). 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 $x \notin (-\sigma, \sigma)$. (**E**) Demonstration of the effect of the scale t on the PHATE visualization for the artificial tree data colored by branch. 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 visualization with different values of t. The red dots in the first column indicate the values of t chosen for the plots. The red dot surrounded by a black box indicate the chosen value of t for the visualization in Figure 1B of the artificial tree data. Values of t that are too low can give noisy visualizations while very high values of t can result in a loss of information in the visualization. (**F**) The PHATE visualization for the iPSC CyTOF dataset from [14] with varying scale parameter t. The embeddings for all t preserve the branching structure and the visualizations are very similar to each other, demonstrating that the embedding is robust to the choice of t. (**G**) The PHATE visualization for the iPSC mass cytometry dataset from [14] with varying number of subsample sizes N. The main branches present for $N = 10000$ are also visible for the other values of N, demonstrating that the PHATE embedding is robust to the size of the subsample.

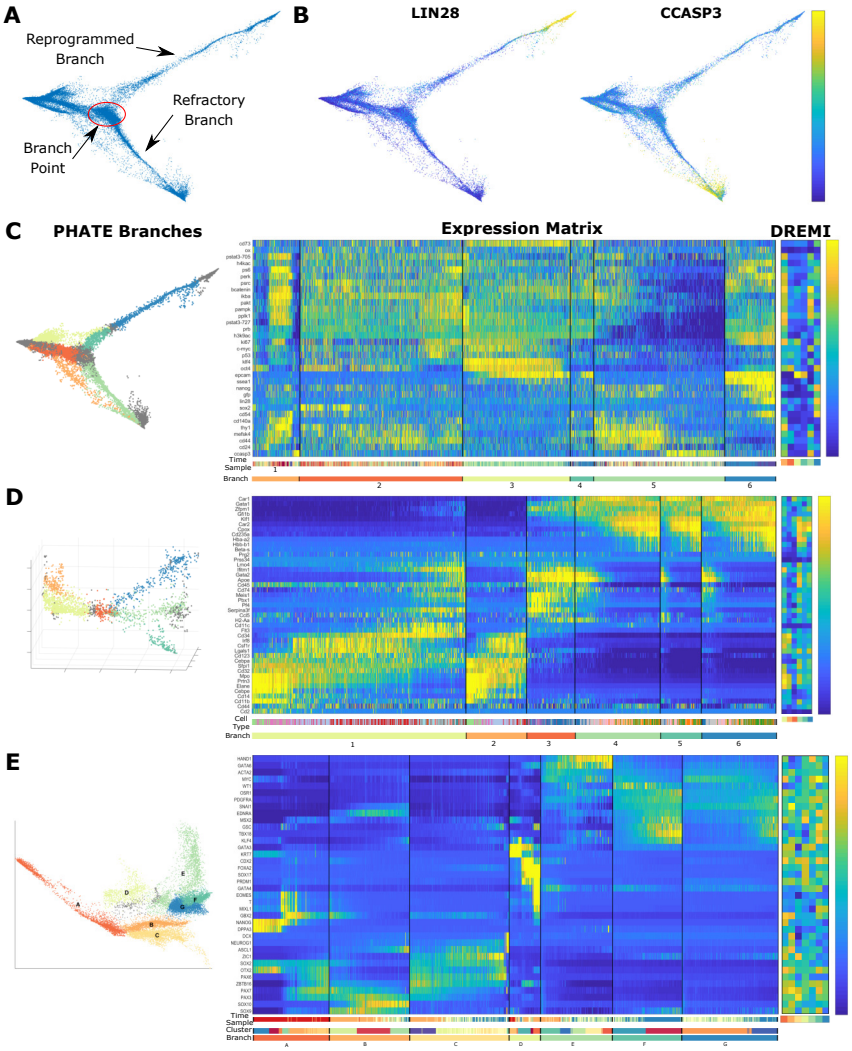
biological data, branch points often encapsulate switch-like decisions where cells sharply veer towards one of a small number of fates. For example, Figure S3A shows PHATE on a CyTOF dataset of induced pluripotent stem cells (iPSC) [14], with a central branch point identified. This branch point connects the early stages of cells with a branch of cells that are successfully reprogrammed and a branch of cells that are refractory (identified via selected markers including those in Figure S3B) and marks a major decision point between these two cell fates. Identifying branch points in biological data is of critical importance for analyzing such decisions.

We make a key observation that most points in PHATE plots of biological data lie on low-dimensional progressions with some noise as demonstrated in Figure 3Aii. Since branch points lie at the intersections of such progressions, they have higher local intrinsic dimensionality. We can also regard intrinsic dimensionality in terms of degrees of freedom in the progression modeled by PHATE. If there is only one fate possible for a cell (i.e. a cell lies on a branch as in Figure 3Aii) then there are only two directions of transition between data points—forward or backward—and the local intrinsic dimension is low. If on the other hand, there are multiple fates possible, then there are at least three directions of transition possible—a single direction backwards and at least two forward. This cannot be captured by a one dimensional curve and will require a higher dimensional structure such as a plane, as shown in Figure 3Aii. Thus, we can use the concept of local intrinsic dimensionality for identifying branch points.

We use a k-nn based method for estimating local intrinsic dimensionality [21]. This method uses the relationship between the radius and volume of a d-dimensional ball. The volume increases exponentially with the dimensionality of the data. So as the radius increases by δ, the volume increases by δ^d where d is the dimensionality of the data. Thus the intrinsic dimension can be estimated via the growth rate of a k-nn ball with radius equal to the k-nn distance of a point. For more details o this approach, see Methods. We note that other local intrinsic dimension estimation methods could be used such as the maximum likelihood estimator in [22]. Figure 3Aii shows that points of intersection in the artificial tree data indeed have higher local intrinsic dimensionality than points on branches.

Endpoint Identification We also identify endpoints in the PHATE embedding. These points can correspond to the beginning or end-states of differentiation processes. For example, Figure S3A shows the PHATE visualization of the iPSC CyTOF dataset from [14] with highlighted endpoints, or end-states, of the reprogrammed and refractory branches. While many major endpoints can be identified by inspecting the PHATE visualization, we provide a method for identifying other endpoints or end-states that may be present in the higher dimensional PHATE embedding. We identify these states using data point centrality and distinctness as described below.

First, we compute the centrality of a data point by quantifying the impact of its removal on the connectivity of the graph representation of the data (as defined using the local affinity matrix $K_{k,\alpha}$). Removing a point that is on a one dimensional progression pathway, either branching point or not, breaks the graph into multiple parts and reduces the overall connectivity. However, removing an endpoint does not result in any breaks in the graph. Therefore we expect endpoints



Supplemental Figure S3: Annotated PHATE visualizations of CyTOF iPSC data [14] and branch expression analysis. (**A**) The primary branch point between the two major branches (reprogrammed and refractory) of the data is highlighted. (**B**) The PHATE visualization colored by LIN28 (a marker associated with the transition to pluripotency [15]) and CCASP3 (associated with cell apoptosis). LIN28 expression is limited to the reprogrammed branch while CCASP3 is primarily expressed in the refractory branch, indicating that the failure to reprogram may initiate apoptosis in these cells. (**C**) Analysis of branches on the PHATE embedding for the iPSC CyTOF data [14], (**D**) bone marrow scRNA-seq dataset from [16], and (**E**) newly generated embryoid body scRNA-seq data. (Left) The PHATE visualization with identified branches. (Middle) Expression level for each cell ordered by branch and ordering within the branch. Cell ordering is calculated using Wanderlust [17] starting on the left-most point of each branch. Expression levels are z-scored for each gene. A colorbar is given below the expression matrices that identifies each branch and (in the case of the bone marrow scRNA-seq data) cell type. (Right) DREMI scores [18] between gene expression levels and cell order within each branch. MAGIC [19] is applied first in (**D**) and (**E**) to impute missing values using the same kernel used for PHATE and smaller t. For branch analysis of the bone marrow data in (**D**), we used 3 PHATE dimensions to obtain clearer branch separation.

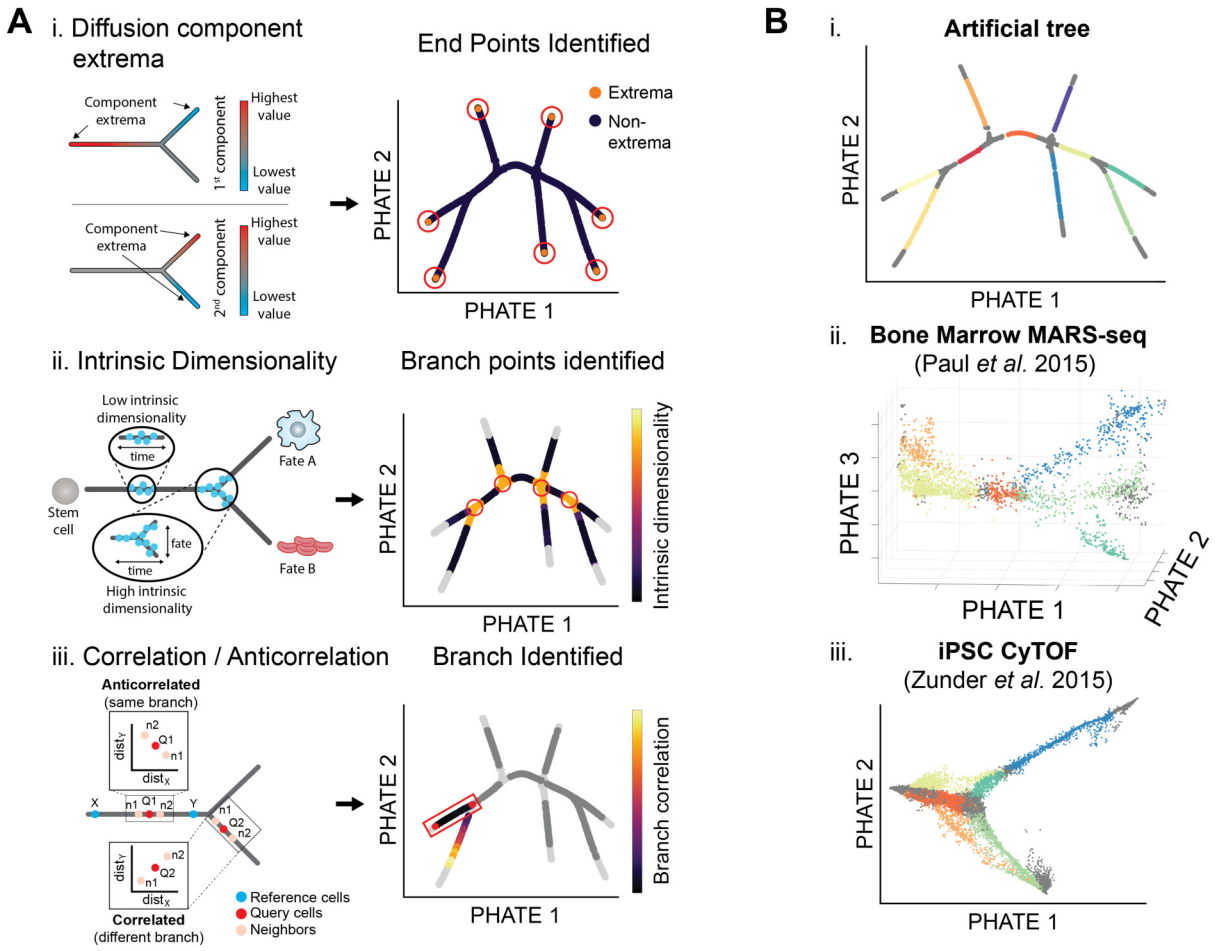


Figure 3: Extracting branches and branchpoints from PHATE. (**A**) Methods for identifying suggested endpoints, branch points, and branches. (i) PHATE computes a specialized diffusion operator as an intermediate step (Figure 2D). We use this diffusion operator to find endpoints. Specifically we use the the extrema of the corresponding diffusion components (eigenvectors of the diffusion operator) to identify endpoints [20]. (ii) Local intrinsic dimensionality is used to find branchpoints in a PHATE visual. As there are more degrees of freedom at branch points, the local intrinsic dimension is higher than through the rest of a branch. (iii) Cells in the PHATE embedding can be assigned to branches by considering the correlation between distances of neighbors to reference cells (e.g. branch points or endpoints). (**B**) Detected branches in the (i) artificial tree data, (ii) bone marrow scRNA-seq data from [16], and (iii) iPSC CyTOF data from [14].

to have low centrality, as estimated using the eigenvector centrality measure of $K_{k,\alpha}$. For more details see Methods.

Second, we quantify the distinctness of a cellular state relative to the general data. We expect the beginning or end-states of differentiation processes to have the most distinctive cellular profiles. As shown in [20] we quantify this distinctness by considering the minima and the maxima of diffusion eigenvectors (see Figure 3Ai). Thus we identify endpoints in the embedding as those that are most distinct and least central.

Branch Identification After identifying branch points and endpoints, the remaining points can be assigned to branches between two branch points or between a branch point and endpoint. Due to the smoothly-varying nature of centrality and local intrinsic dimension, the previously described procedures identify regions of points as branch points or endpoints rather than individual points. However, it can be useful to reduce these regions to representative points for analysis such as branch detection and cell ordering. To do this, we reduce these regions to representative points using a "shake and bake" procedure similar to that in [23]. Essentially, this approach groups collections of branch points or endpoints together into representative points based on their proximity (see Methods for details). Further, a representative point is labeled an endpoint if the corresponding collection of points contains one or more endpoints as identified using centrality and distinctness. Otherwise, the representative point is labeled a branch point.

After representative points have been selected, the remaining points can be assigned to corresponding branches. We use an approach based on the branch point detection method in [3] that compares the correlation and anticorrelation of neighborhood distances. However, we use higher dimensional PHATE coordinates instead of the diffusion maps coordinates. Figure 3Aiii gives a visual demonstration of this approach and details are given in Methods. Figure 3B shows the results of our approach to identifying branch points, endpoints, and branches on an artificial tree dataset, a scRNA-seq dataset of the bone marrow [16], and an iPSC CyTOF dataset [14]. Our procedure identifies the branches on the artificial tree perfectly and defines biologically meaningful branches on the other two datasets which we will use for data exploration.

2.3 Comparison of PHATE to Other Methods

To demonstrate the effectiveness of PHATE, we extensively compare PHATE to other methods on synthetic and biological datasets with different types of structure. First, we compare PHATE to other methods on artificial testcases (see Figure 4) that we generate with known ground truth (shown in the panel on the left in Figure 4).

The first artificial dataset is a tree with seven branches and three intermediate segments (red, orange and green) in high dimensions, shown with two different levels of noise (see Methods for details). This dataset tests the ability of embedding methods to reveal branching trajectory structures that may commonly exist in differentiation systems. Of the methods shown, only PHATE shows the correct number of branches and connections at both levels of noise. t-SNE and UMAP tend to shatter the embedding in low noise conditions. Diffusion maps oversimplify

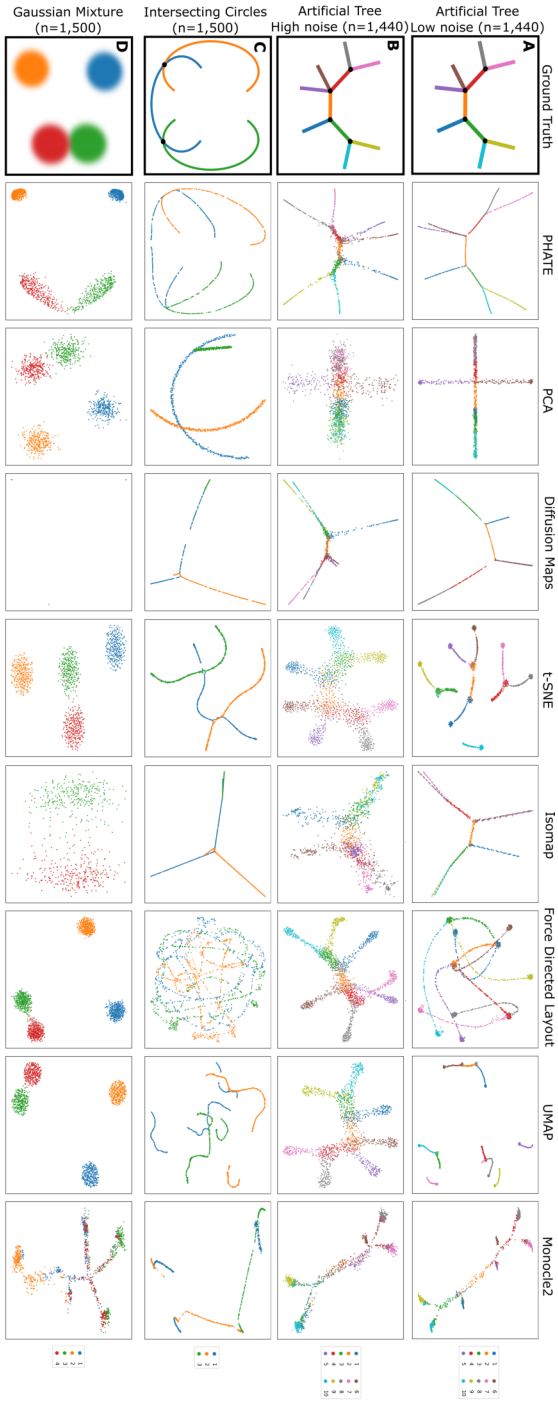


Figure 4: Comparison of PHATE to other visualization methods on artificial datasets containing clusters, trajectories, and intersecting manifolds. Ground truth based on the generative process of each dataset is shown in the leftmost panels. Columns represent different visualization methods, rows different datasets.

the branching structure, as each branch is generally shown in one diffusion component. Monocle2 fails to detect some branches in high noise, and all other methods fail in both low and high noise conditions. Failure here is defined as the inability to show the correct number of branches or correct connections.

The second dataset consists of three intersecting circles (orange, blue, and green). This dataset is designed to test the ability of different methods to handle manifold-intersections and boundary-conditions. Of the methods shown, only PHATE shows the right number of circles and correct intersections. Diffusion maps show instability at intersections as is further explored in Supplemental Figure S1 Column 4. Isomap and Monocle2 also show simplified structure, t-SNE and UMAP once again lead to shattered embeddings and other methods show generally noisy unintelligible embeddings.

We also show the methods on data with a Gaussian Mixture Model. Here we see that while several methods show cluster structures, PHATE more explicitly shows the relationships between clusters. Additionally, t-SNE, Force Directed Layout and UMAP fail to retain the correct global relationships between distant clusters: the ground truth shows blue adjacent to green and orange adjacent to red; t-SNE and Force Directed Layout swap these while UMAP distorts the spatial relationships entirely. By contrast Monocle2 still shows a tree-structure, and DMs and ISOMAP miss most of the structure. Since clusters, intersecting manifolds and branching trajectories are common structures in biological datasets, we expect similar performance from the methods on real datasets featuring these structures.

Next, we compare each of the visualization methods on five biological datasets containing a combination of clusters and branching trajectories (see Figure 5).

The first biological dataset is scRNA-seq data measured from mouse bone marrow cells enriched for myeloid and erythroid lineages. The embeddings are colored by the clusters reported in [16]. In Section 3, we show that PHATE reveals the structure of this data as a primary progression with a trifurcation in the erythrocyte cells. This contrasts with the results presented in [16] which focused on cluster structure. At the same time, PHATE maintains cluster coherence (i.e., cells in the same cluster are generally close together in the visualization). DMs shows a highly simplified structure, and Monocle2 shows a four-branched tree which fails to maintain cluster coherence. Multiple other methods also show some level of branching structure (e.g., Force Directed Layout), indicating that the branching structure detected by PHATE is real.

The second biological dataset is scRNA-seq data measuring retinal bipolar neurons. The embeddings are colored by the cluster assignments reported in [13]. Cells were collected from an adult mouse and sorted for transgenic retinal bipolar markers. PHATE visualizes cluster structure while preserving relationships between clusters. DMs here are strongly affected by outliers, PCA and Isomap merge many of the smaller clusters, Monocle2 forms a tree which has seemingly no correspondence with clusters, and t-SNE fails to preserve the distance between rod bipolar cells (orange) and Müller glia (red) from the remaining clusters (mostly cone bipolar cells.) We further show in Figure 5 that PHATE's separation of the rod bipolar (RBC) cluster into a bifurcating trajectory is consistent with morphological observations [24] indicating that RBCs can be separated into at least three subclasses. The third biological dataset is

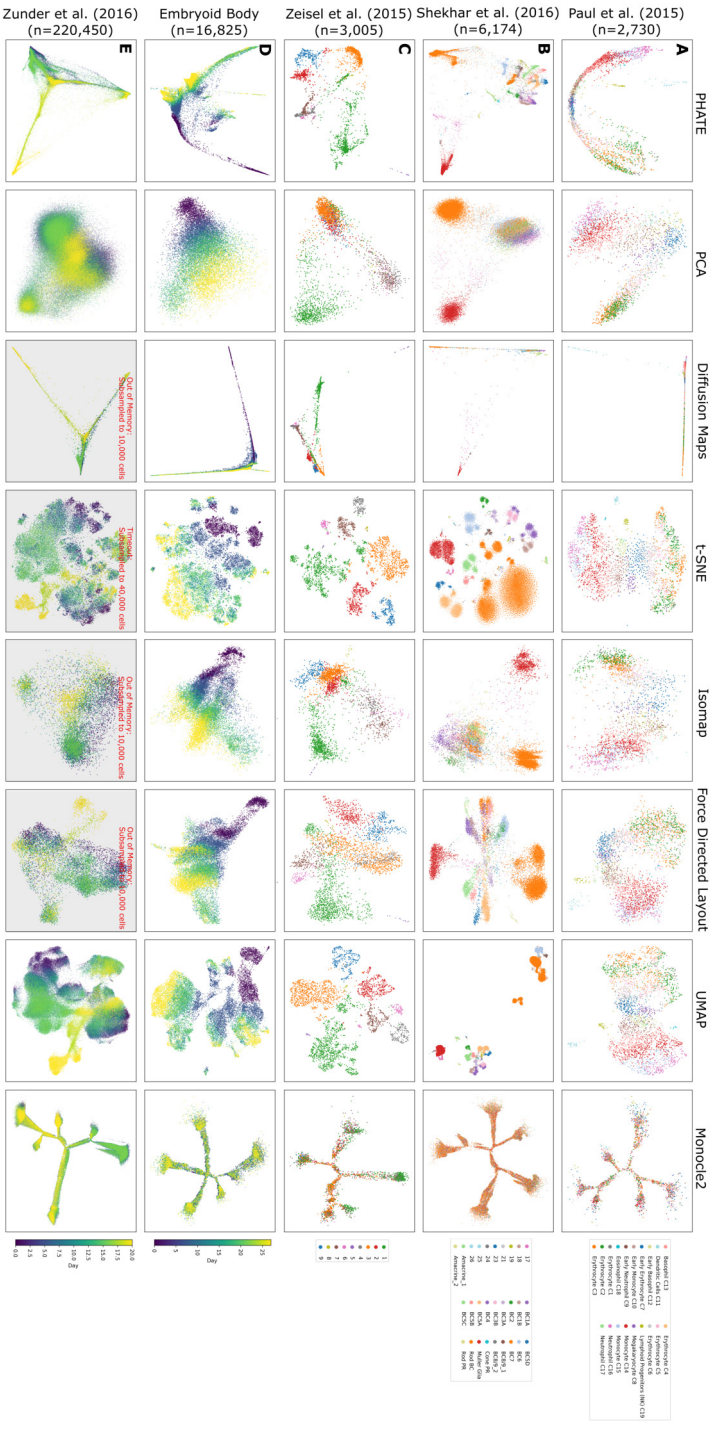


Figure 5: Comparison of PHATE to other visualization methods on biological datasets. Columns represent different visualization methods, rows different datasets.

scRNA-seq data obtained from cells in the mouse cortex and hippocampus [25]. Points are colored by cluster assignments obtained using Phenograph [26]. PCA, Isomap and Force Directed Layout all fail to separate some of the clusters, while t-SNE and UMAP shatter cluster 1 (green). Monocle2 creates a spurious trajectory structure which fails to separate the clusters. Thus we are showing that PHATE maintains cluster separations while often inferring some continuity or progression within the clusters.

The fourth biological dataset is scRNA-seq data from human embryonic stem cells (hESCs) differentiating in an embryoid body (EB) setting, that we collected at 3-day intervals over a 27-day differentiation time course and colored by time point. PCA and Isomap give amorphous structures that preserve the overall time progression of the data while failing to separate multiple lineages. DMs gives a highly simplified trajectory that is missing multiple branches such as the developing neural crest. UMAP and t-SNE shatter the trajectories into distinct clusters, even destroying the natural time progression within the data (between early embryonic stem cells seen in earlier days, colored in red and lineages that emerge in later days colored in blue). Monocle2 forms a tree whose branches are not correlated with time. Monocle2 also fits a different tree to the data each time it is run, making it difficult to determine the true structure of the data. Only PHATE and Force Directed Layout give a combination of clusters and trajectories which retains distinct time points and recapitulates the biological interpretation explored in detail in Section 4. However, PHATE provides a much cleaner and denoised visualization than Force Directed Layout, which tends to splay out points.

The final biological dataset we use in our comparisons is a mass cytometry measurement of cellular reprogramming of mouse embryonic fibroblasts (MEFs) to induced pluripotent stem cells (iPSCs) from [14], colored by time point. Here, PCA, t-SNE, Isomap, Force Directed Layout, and UMAP all fail to show the branching structure described in Section 3. DMs display a branching structure, but several branches that represent intermediate, partially reprogrammed states are not present. Monocle2 generates a tree with more branches than are visible on the PHATE embedding; However, since Monocle2 generates a different structure each run, the validity of these branches is difficult to ascertain.

Further comparisons are shown in Figure S1 with additional datasets and methods, including MDS, Locally Linear Embedding, MDS on Diffusion Maps, and t-SNE on Diffusion Maps. Together, all of these comparisons highlight that each step in PHATE is necessary to achieve a useful visualization. In particular, the comparisons to MDS or t-SNE on Diffusion Maps (neither of which produce useful visualizations) demonstrate the necessity of the informational distance embedding for a stable and useful visualization. t-SNE on Diffusion Maps essentially produces a denoised version of t-SNE, failing to preserve the global distances encoded in the diffusion operator due to the nature of the local neighborhood penalty in t-SNE, and failing to handle boundary conditions due to the shortcomings of Diffusion Maps. MDS on Diffusion Maps suffers from the same issues that affect Diffusion Maps, failing to handle boundary conditions and intersecting manifolds and collapsing local structure. Hence, the additional steps taken in PHATE to transform the diffusion operator into an embedding amenable to 2D/3D visualization are necessary for the successful visualization of data-driven diffusion geometry.

Table S1 provides a summary of the attributes of the methods shown in Figure S1 for further comparison.

3 PHATE for Data Exploration

PHATE reveals the underlying structure of the data. In this section, we show the insights gained through the PHATE visualization of this structure for primarily single-cell data. Section 6.3 in the Supplementary Text discusses PHATE applied to more general datasets, including SNP data, microbiome data, Facebook network data, Hi-C chromatin conformation data, and facial images (Figure 6).

We show that the identifiable trajectories in the PHATE visualization have biological meaning that can be discerned from the gene expression patterns and the mutual information between gene expression and the ordering of cells along the trajectories. We analyze the mouse bone marrow scRNA-seq [16] and iPSC CyTOF [14] datasets described previously. For both of these datasets, we used our new methods for detecting branches and branch points. We then ordered the cells within each trajectory using Wanderlust [17] applied to higher-dimensional PHATE coordinates. Ordering is generally from left to right. We note that ordering could also be based on other pseudotime ordering software such as those in [2] or [3]. To estimate the strength of the relationship between gene expression and cell ordering along branches, we estimated the DREMI score (a conditional-density resampled mutual information that eliminates biases to reveal shape-agnostic relationships between two variables [18]) between gene expression and the Wanderlust-based ordering within each branch. Genes with a high DREMI score within a branch are changing along the branch.

iPSC Mass Cytometry Data Figure S3C shows the mass cytometry dataset from [14] 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, further segmented in our visualization, each corresponding to biology identified in [14]. Branch 2 contains early reprogramming intermediates with the correct set of reprogramming factors $Sox2^+/Oct4^+/Klf4^+/Nanog^+$ and with relatively low CD73 at the beginning of the branch. Branch 2 splits into two additional branches. Branches 4 and 6 (Figure S3) show the successful reprogramming to ESC-like lineages expressing markers such as Nanog, Oct4, Lin28 and Ssea1, and Epcam that are associated with transition to pluripotency [15]. Branch 5 shows a lineage that is refractory to reprogramming, does not express pluripotency markers, and is referred to as "mesoderm-like" in [14].

Then, branch 3 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 Epcam, which is known to promote reprogramming

	Denoises	Preserves local structure	Preserves global structure	Metric embedding	Non-linear	Handles boundary conditions	Isometrically stable embedding	Practically scalable	Optimized for 2D/3D visualization	Structural Assumptions
PHATE	✓	✓	✓	✓	✓	✓	✓	✓	✓	Manifolds
PCA	✗	✗	✓	✓	✗	✓	✓	✓	✗	Linear
DM	✓	✓	✓	✓	✓	✗	✓	✓	✗	Manifolds
MDS	✗	✗	✗	✓	✓	✓	✗	✗	✓	None
MDS on DM	✓	✗	✗	✓	✓	✗	✗	✗	✓	Manifolds
t-SNE	✗	✓	✗	✗	✓	✓	✗	✗	✓	Clusters
t-SNE on DM	✓	✓	✗	✗	✓	✗	✗	✗	✓	Clusters
LLE	✓	✓	✗	✓	✓	✗	✓	✗	✗	Manifolds
Isomap	✗	✓	✗	✓	✓	✓	✓	✗	✗	Manifolds
FDL	✓	✓	✗	✗	✓	✓	✗	✗	✓	Graph
UMAP	✗	✓	✓	✗	✓	✓	✓	✓	✓	Clusters
Monocle2	✓	✓	✓	✗	✗	✓	✗	✓	✓	Tree

Supplemental Table S1: Comparison of 12 visualization methods across a series of desirable criteria. Metric embedding: euclidean distances in the embedding have metric interpretation in the ambient space. Isometrically stable embedding: the embedding is consistent (excluding rotation and scale) when run multiple times on the same dataset. Practically scalable: can be run on most single-cell datasets ($< 250,000$ cells) with 128GB of RAM in less than one hour. Optimized for 2D/3D visualization: the number of components desired affects the nature of the embedding such that a 2D embedding contains more information than the first two dimensions of an n dimensional embedding.

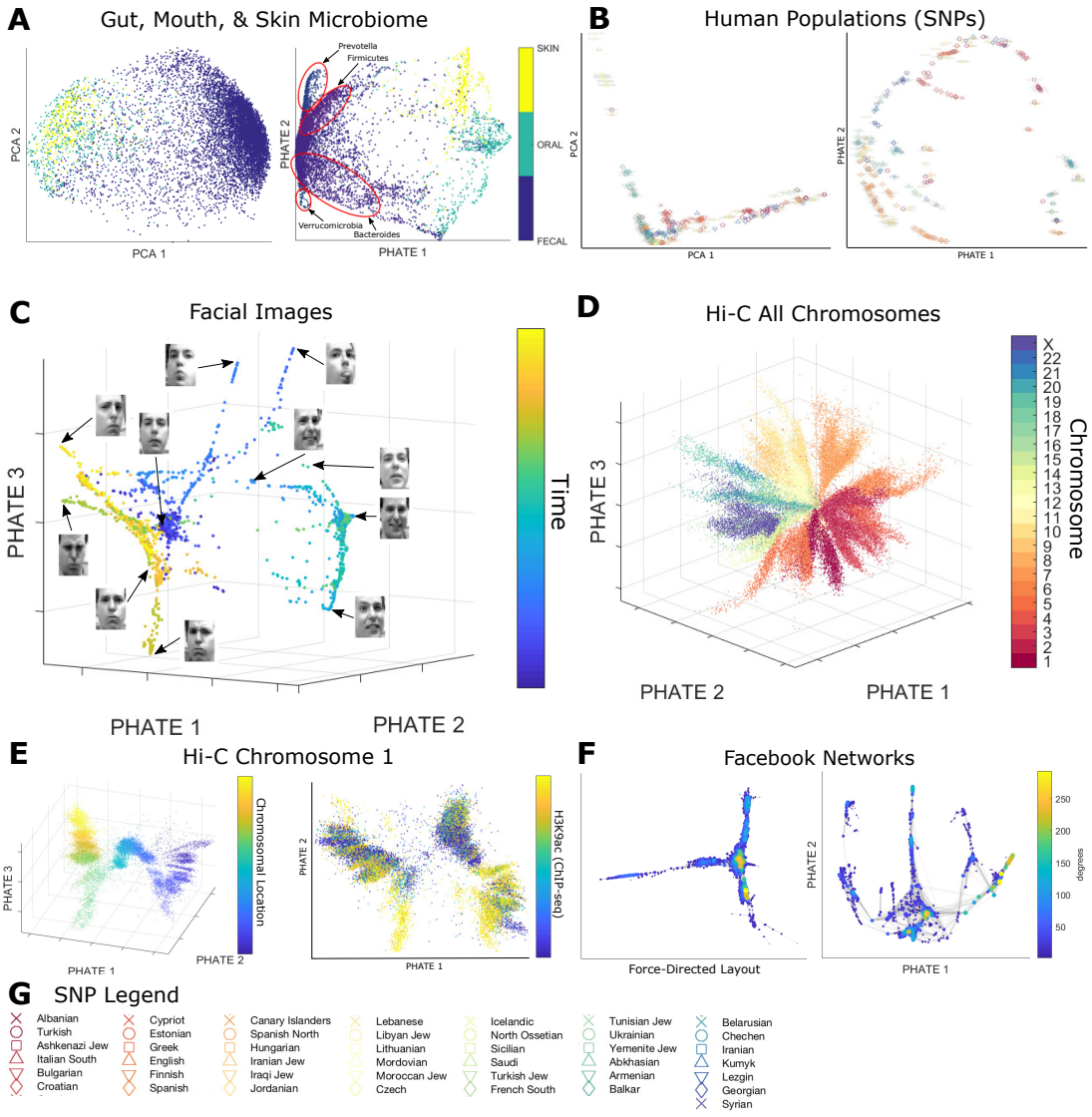


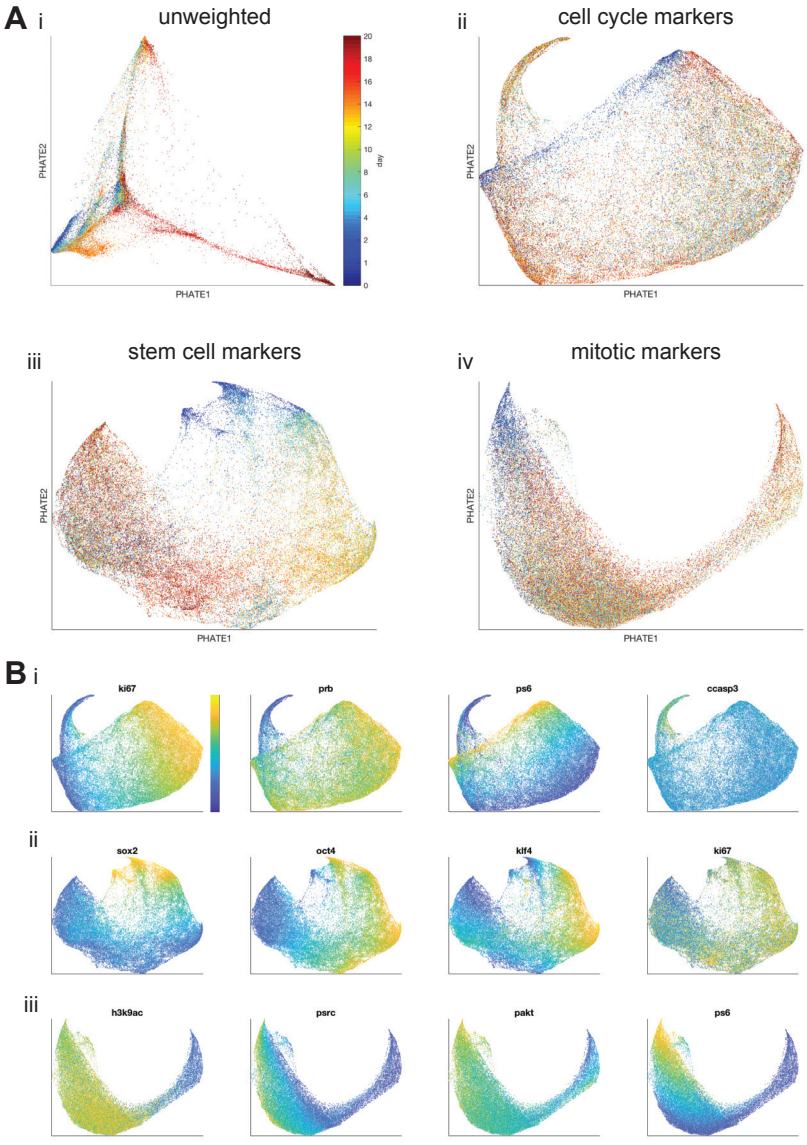
Figure 6: PHATE reveals structure in a variety of datasets. (**A**) PCA and PHATE embeddings of microbiome data from the American Gut project, colored by body site, and branches annotated by their dominant genera or phyla (Figure S6). (**B**) PCA and PHATE embeddings of SNP data from the Human Origins dataset showing genotyped present-day humans from 203 populations [27]. (**C**) A 3D PHATE visualization of the Frey Faces dataset used in [28]. Points are colored by time within the video. Multiple branches corresponding to different poses are clearly visible. (**D**) 3D PHATE visualization of human Hi-C data [29] using all 23 chromosomes at 50 kb resolution, colored by chromosome. Each point corresponds to a genomic fragment. (**E**) PHATE visualizations of the same human Hi-C data in **B** for chromosome 1 at 10 kb resolution colored by chromosome location (left) and H3K9ac ChIP-seq signal (right). (**F**) Force-directed layout and PHATE visualizations of Facebook network data with data points colored by their degree (number of connections). (**G**) Population legend for the SNP data in **B**.

generally [30], increases along this branch. This is evidenced by the high DREMI score between Epcam and the cell ordering within the branch (Figure S3C). This branch joins into branch 4 at a later stage, showing perhaps an alternative path or timing of reprogramming. Finally, branch 1 shows a lineage that has failed to reprogram, perhaps due to the wrong stoichiometry of the reprogramming factors [31]. Of note, this lineage contains low Klf4 which is an essential reprogramming factor.

Additionally, the PHATE embedding shows a decrease in p53 expression in precursor branches (2 and 3) indicating that these cells are released from cell cycle arrest induced by initial reprogramming factor over expression [32]. However, along the refractory branch (branch 5) we see an increase in cleaved-caspase3, potentially indicating that the failure to reprogram correctly initiates apoptosis in these cells [14].

PHATE on different views of the data By default PHATE produces a single low dimensional embedding of a dataset. However, we can obtain variants of this embedding by reweighting the features before computing distances. Such reweightings correspond to specific "views" of the data. For example, in a biological context, we can upweight genes that are involved in a specific process in order to have PHATE prominently reflect this process. To demonstrate this reweighting scheme, we computed three alternative PHATE embeddings of the iPSC data, by upweighting either cell cycle markers, stem cell markers, or mitotic markers (Figure S4). PHATE, after upweighting cell cycle markers, gives an embedding with a circular structure (Figure S4Aii) that reflects the cyclical nature of the cell cycle. In addition to the circular structure, the embedding shows a small protrusion, with high expression of Ccasp3, suggesting that these cells are apoptotic. Upweighting stem cell markers gives an embedding with a 1-dimensional progression. Expression analysis reveals that stem cell markers such as Sox2 are high at one end of the progression and low on the other end. Moreover, the progression is correlated with time (measurement day), further supporting the idea that the progression that PHATE reveals marks the extent to which the cells are stem-like, with early timepoints being less stem-like. Finally, after upweighting mitotic markers, PHATE shows a different 1-dimensional progression. Here, the progression appears to be correlated with mitotic state, as can be seen by the expression of several mitosis-related genes (Figure S4Biii), such as pAKT, that are high only in one end of the embedding. Thus, PHATE computed after reweighting the genes can be used to obtain a process specific embedding to gain insight into predefined biological processes.

Bone Marrow scRNA-seq Data Reveals New Structure Figure S3D shows the color-coded 3D PHATE embedding and gene expression matrix for scRNA-seq data from mouse bone marrow. This data is enriched for myeloid and erythroid lineages and was organized into clusters in [16], as shown in Figure 5A. 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 in the center to erythroid lineages towards the right and myeloid lineages towards the left. The expression matrix shows increasing expression of erythroid markers in the rightmost branches (branches 4,



Supplemental Figure S4: PHATE using reweighted distances to highlight specific biological processes or "views" of the data. (**A**) PHATE embedding of CyTOF iPSC data using (i) unweighted distances, (ii) distances after upweighting cell cycle markers, (iii) distances after upweighting stem cell markers, (iv) distances after upweighting mitotic markers. (**B**) PHATE colored by different markers (columns). From top to bottom: (i) PHATE cell cycle "view", (ii) PHATE stem cell "view" (iii) PHATE mitotic "view".

5, and 6) such as hemoglobin subunits Hba-a2 and Hbb-b1 as well as heme synthesis pathway enzyme Cpox as the lineage progresses to the right. Towards the left in branches 1 and 2, we see an enrichment for myeloid markers, including CD14 and Elane, which are primarily monocyte and neutrophil markers, respectively.

In addition, PHATE splits the erythrocytes into three branches not distinguished by the authors of [16]. These branches show differential expression of several genes. Branch 6 is more highly expressed in Gata1 and Gfi1B, both of which are involved in erythrocyte maturation. Branch 4 is also more highly expressed in Zfpm1 which is involved in erythroid and megakaryocytic cell differentiation. Additionally, branches 4 and 5 are more highly expressed in Car2, which is associated with the release of oxygen. Given these differential expression levels, it is likely that the different branches correspond to erythrocytes at different levels of maturity and in different states [33–39]. In addition, the branches at the right have high mutual information with CD235a, which is an erythroid marker that progressively increases in those lineages.

PHATE in 3 dimensions more clearly reveals a separation of the myeloid lineages to the left into two branches. Cd14 and Sfpi1 are both more highly expressed at the beginning of branch 2 than in branch 1, suggesting that branch 2 is associated with monocytes while branch 1 is associated with neutrophils.

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 Figure 4D). However, we note that both the t-SNE and PCA embeddings of this data in Figure S1 also lack a gap between these trajectories.

PHATE reveals transcriptional heterogeneity in Rod Bipolar Cells Figure S5Fi shows PHATE on scRNA-seq data of mouse retinal bipolar neurons from [13]. Cells were collected from an adult mouse and sorted for transgenic retinal bipolar markers. PHATE visualizes cluster structure while preserving relationships between clusters. The embedding is colored by the clusters described in the original study, which seeks to transcriptionally characterize all subtypes of bipolar cells. In the original characterization, rods bipolar cells (the largest cluster of cells) are shown as a single homogeneous cell type. However, the PHATE embedding in Figure S5Fi reveals a bifurcating trajectory within this cluster. We zoomed in on this bifurcating trajectory, by embedding just those cells, in order to determine their sub-structure.

This embedding of just rod bipolar cells (Figure S5Fii) reveals four distinct sub-clusters (found by k-means clustering) of rod bipolar cells. We characterize the transcriptional profile of these sub-clusters in Figure S5G, showing all genes used for cell type assignment in [13]. Our results show a trajectory between rod bipolar cell types that is consistent with previous work by [24], in which cell types RB1 and RB2 are shown to be a continuum of variants of a single type. Further, we show distinct differences between these in known marker genes distinguishing RB1 and RB2 *Trnp1*, *Rho* and *Pde6b* [40], indicating that the four clusters we observe may be further subtypes of these two cell types.

PHATE on Large Scale Datasets To demonstrate the scalability of PHATE for data exploration on large datasets, we applied PHATE to the 1.3 million mouse brain cell dataset from 10x [41]. Figure S5C shows a comparison of PHATE to t-SNE, colored by 10 of the 60 clusters provided by 10x. We see that PHATE retains cluster coherence while t-SNE shatters some of the cluster structure.

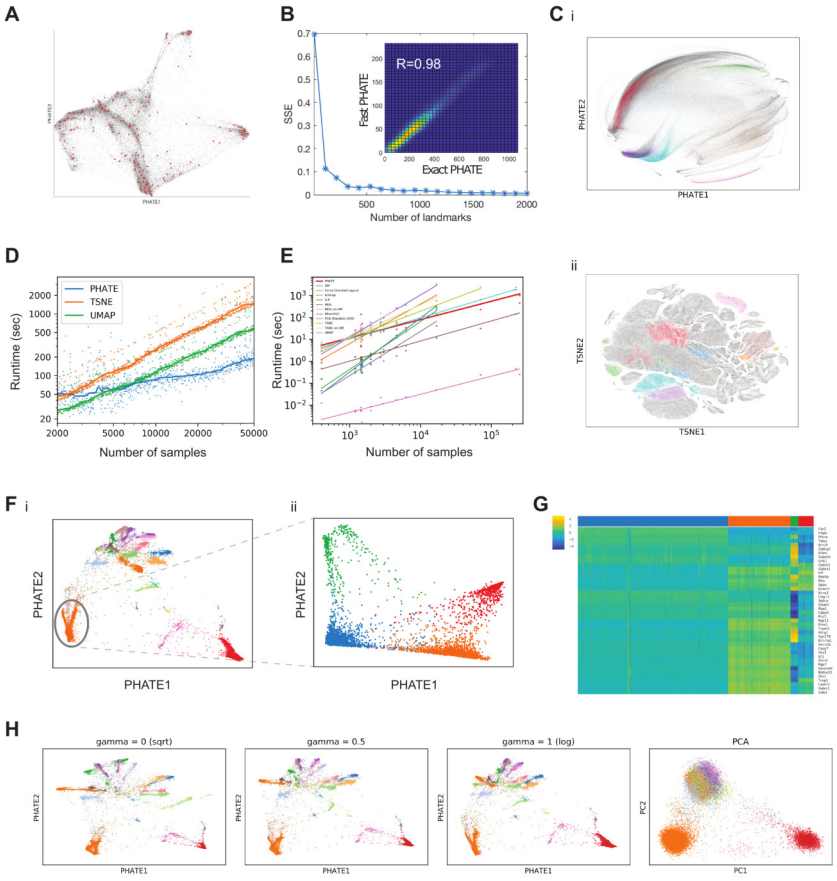
We also ran PHATE on a network dataset of 1.8 million Wikipedia articles and 29 million hyperlinks between them [42]. PHATE took approximately 12 minutes to compute the embedding, greatly outperforming any existing graph layout methods such as force-directed layout which could not be run on the entire dataset. To show that the embedding obtained by PHATE is meaningful, we colored a subset of the points by selected Wikipedia category (Figure S6E). PHATE maintains coherence between article topics.

4 Exploratory Analysis with PHATE on Human ESC Differentiation Data

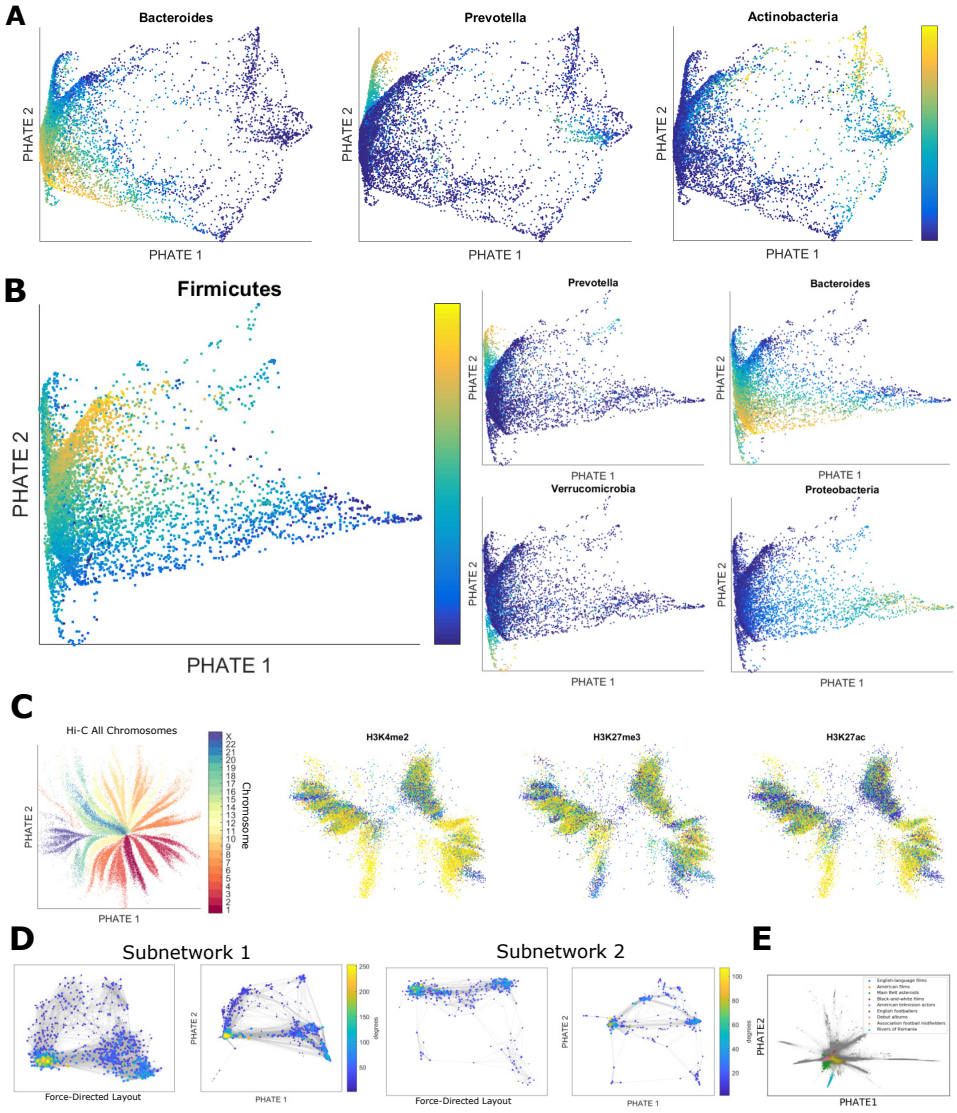
To validate the ability of PHATE to reveal biological insights in newly measured systems, we generated scRNA-seq data from human embryonic stem cells (hESCs) differentiating in an embryoid body (EB) setting [43]. We show that 1. the PHATE embedding corroborates with known biology, 2. it can be used to find features for experimental isolation (FACS sorting) of subpopulations and lineages of interest to biologists, 3. it can be used to predict transcription factors associated with such lineages, many of which are uncharacterized, and 4. it can be used to understand larger patterns of gene expression through the whole space.

EB differentiation is a multi-step process that begins with the induction of primary germ layers: the ectoderm, endoderm and mesoderm. With time, these germ layer precursors give rise to a diverse array of differentiated cell types. EB differentiation is thought to resemble embryonic development *in vivo* and has been successfully used to produce various types of neurons, astrocytes and oligodendrocytes [44–47], hematopoietic, endothelial and muscle cells [48–56], hepatocytes and pancreatic cells [57,58], as well as germ cells [59,60]. However, the molecular pathways regulating germ layer development are largely unknown. We applied PHATE to the new EB scRNA-seq data to elucidate differentiation trajectories and gene-gene interactions that underlie lineage development.

We measured 31,000 cells equally distributed over a 27-day differentiation time course. Samples were collected at 3-day intervals and pooled for measurement on the 10x Chromium platform (see experimental methods for more details). Although time was not included in creating the embedding, the PHATE visualization of this data captured a strong time trend within the data and revealed greater phenotypic diversity at later differentiation time points (Figure 7A, left).



Supplemental Figure S5: (**A**) Scalable PHATE embedding of iPSC CyTOF data with a subset of the landmarks shown in red (200 out of 2000). (**B**) Robustness of PHATE to the number of landmarks chosen. PHATE on the EB data computed using increasing numbers of landmarks (X-axis) was compared to exact PHATE, i.e. without landmarks. Comparison was done using Procrustes analysis (optimal linear transformation) and the sum of squared error (SSE, Y-axis) is shown. To ensure a stable embedding that accurately approximates exact PHATE we choose 2000 landmarks as default. The inset shows the histogram of pairwise distances in the visualization computed using fast PHATE (2000 landmarks) on the EB data vs. the pairwise distances from exact PHATE. The correspondence and correlation coefficient are very high. (**C**) PHATE and t-SNE embeddings of a 1.3 million mouse brain cell dataset from 10X genomics [41]. PHATE embedding was performed with 2000 landmarks and completed in three hours. A subset (10 of 60) of the clusters provided by 10X are shown in color, the rest in gray. t-SNE shatters the cluster structure, while PHATE retains clusters as contiguous groups of cells. (**D**) Runtime of PHATE, t-SNE and UMAP on increasingly large subsamples of the EB data. Runtime was averaged across four runs. (**E**) Runtime of 12 visualization methods shown in Figure S1 across all 19 displayed datasets and corresponding line of best fit for each method. Where a method ran out of memory or took longer than one hour, the runtime is not shown and linear fits are cut off accordingly. (**F**) i. Initial PHATE embedding of scRNAseq on mouse retinal bipolar neurons. The rod bipolar cells cluster (cluster 1) is circled. ii. Subsequent PHATE embedding of cluster 1, colored by K means clustering to show heterogeneity within rod bipolar cells. (**G**) Transcriptional characterization of subtypes of rod bipolar cells, using known bipolar cell markers. (**H**) Visualization of scRNAseq on mouse retinal bipolar neurons data using different informational distances defined via the parameter γ.



Supplemental Figure S6: PHATE applied to various biological and nonbiological datasets. (**A**) The PHATE embedding of data from the American Gut project colored by 2 genera (bacteroides and prevotella) and a phylum (actinobacteria) of bacteria. (**B**) The PHATE embedding of only the fecal samples from the American Gut project 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. (**C**) 2D PHATE visualization of human Hi-C data [29] for (left) all chromosomes at 50 kb resolution and for chromosome 1 at 10 kb resolution, colored by selected chromatin modification markers. (**D**) Comparison of the force-directed layout and PHATE visualizations of subnetworks within the Facebook network data in Figure 6E. The subnetworks are taken from the friend networks of selected individuals within the entire network. In all cases, PHATE reveals more structure. (**E**) PHATE applied to a network dataset of 1.8 million Wikipedia articles and 29 million hyperlinks between them colored by a subset of Wikipedia categories [42]. PHATE maintains coherence between article topics. Running PHATE on the entire dataset took approximately 12 minutes whereas running force directed layout would take significantly longer.

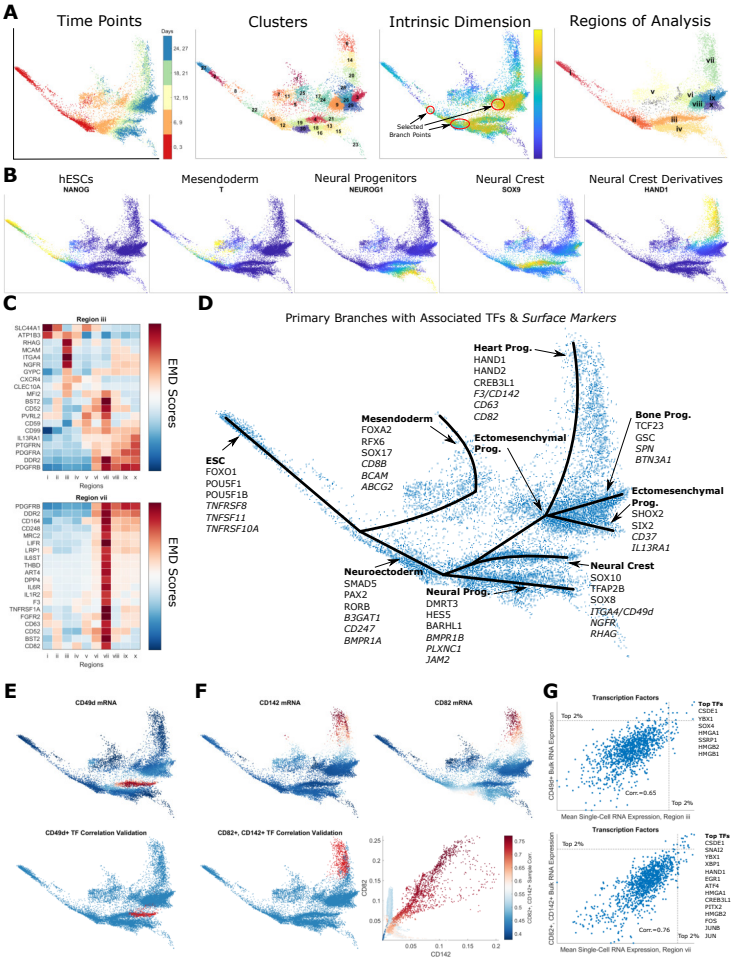


Figure 7: PHATE analysis of embryoid body scRNA-seq data. (**A**) Left: PHATE colored by sample time, i.e., number of days of growth at which the embryoid bodies were measured. Middle left: The PHATE visualization colored by clusters. Clustering is done on a 10-dimensional PHATE embedding. Middle right: The PHATE visualization colored by estimated local intrinsic dimensionality with selected branch points highlighted. Right: Regions of analysis chosen from contiguous clusters. (**B**) PHATE colored by expression levels of selected markers. (**C**) The earth mover distance (EMD) [61] scores of the top scoring surface markers in the targeted regions (regions iii and vii). (**D**) Overlay of the PHATE visualization of the EB scRNA-seq data with branches labeled with associated transcription factors and surface markers (in italics) based on a differential expression analysis using EMD. (**E**) PHATE colored by CD49d expression level from the scRNA-seq data (top) and by correlation between the scRNA-seq transcription factor expression and the CD49d-sorted bulk RNA-seq transcription factor expression per cell (bottom). (**F**) Same as **E**, with CD142 and CD82. The correlation coefficient is highest in region vii, which is the region with the highest CD142 and CD82 expression. Bottom right: Scatter plot of single cell expression levels between CD82 and CD142. Color corresponds to the correlation between the scRNA-seq expression and the CD142+CD82+ sorted bulk RNA-seq expression. The region with highest correlation corresponds to cells that are positive in both CD142 and CD82. (**G**) Scatter plots of the bulk transcription factor expression vs. the mean single-cell transcription factor expression in regions iii (left) and vii (right). The top transcription factors in both the single cell and the bulk data are highlighted.

4.1 Clustering Analysis

We first clustered the data to partition it into regions for analysis (Figure 7A, middle left). Clustering was performed using the k-means algorithm on 10 PHATE dimensions to capture the higher-dimensional structure. This can be viewed as a variant on spectral clustering using PHATE dimensions instead of the eigenvectors of a Laplacian matrix [62, 63]. Clusters were assigned to three major lineages, neuroectoderm, mesendoderm, and neural crest, based on the expression of lineage markers imputed using the MAGIC algorithm [19] (Figures 7B, S3E; Table S2). However, this data set is dominated by continuous progression structure rather than discrete cluster structure. Therefore, we also analyzed the branching progressions.

4.2 Progression Analysis

We identified progressions and branches for analysis by using the same methods we used to analyze the single cell data: The local intrinsic dimensionality was estimated on higher-dimensional PHATE dimensions to identify suggested regions as branch points (Figure 7A, middle right). The centrality measure and diffusion map extrema were used to identify possible end points. We then selected branches by concatenating clusters between the selected branch and endpoints based on their spatial contiguity within the visualization (Figure S3E, left). Finally, Wanderlust [17] was used to define an ordering of cells within each branch where the starting cell of each branch was chosen as the left-most cell.

Figure S3E shows the gene expression matrix within the identified branches for a set of well-characterized lineage-specific markers. From this matrix, we identified different cell types and differentiation processes associated with the branches and clusters. ESC-specific transcripts NANOG and DPPA3 were highly expressed in cells located at the left-most part of branch A (cluster 27), indicating that this is the starting point of the data. As cells travel along branch A through clusters 27, 2, 8, and 22, the epiblast marker OTX2 is upregulated and then downregulated, followed by a sharp increase in MIXL1, EOMES, and T levels in cluster 22, indicating that mesendoderm differentiation begins in this region. These mesendoderm markers continue to be highly expressed in cells located in branch D in cluster 7, followed by high expression of the definitive endoderm markers FOXA2 and SOX17 throughout branch D.

Further along branch A, past the mesendoderm initiation region, the neuroectoderm/early neural crest markers PAX6, ZBTB16, GBX2, PAX3, and PAX7 are induced in clusters 10 and 12. These early progenitors further resolve into neuronal and late neural crest lineages with characteristic markers expressed along each branch. Branch B is characterized by high expression of PAX3, PAX7, SOX9, and SOX10, which are all associated with neural crest differentiation. In contrast, branch C shows high expression of the neural progenitor markers ASCL1, ZIC1, SOX2, NEUROG1, and DCX. Branches E, F, and G are enriched in genes expressed in ectomesenchymal neural crest derivatives including cartilage, bone, smooth muscle, and adipocyte progenitors (SNAI1, TWIST, WT1, OSR1, PDGFRA, TBX18, ACTA2, GSC, KLF4). Figure 7B shows the PHATE visualization colored by selected genes to highlight the different

regions. This analysis demonstrates that the PHATE embedding can successfully resolve germ layers during in vitro differentiation of ESCs.

4.3 PHATE Corroborates Known Biology

To show that the PHATE embedding agrees with known biology, we compared the scRNA-seq data with bulk RNA-seq data available from the literature including datasets from ESCs [64]; ESC-derived neuroectodermal cells (NEC) [65]; ESC differentiating into neural progenitor cells (NPC) at days 0, 8, and 16 [66]; neural crest cells (NCC) [67]; definitive endoderm cells (DEC) [64]; dental pulp stem cells (DPSC, derived from neural crest cells) [68]; human foreskin fibroblasts (HFF) [64]; and cardiomyocytes [69].

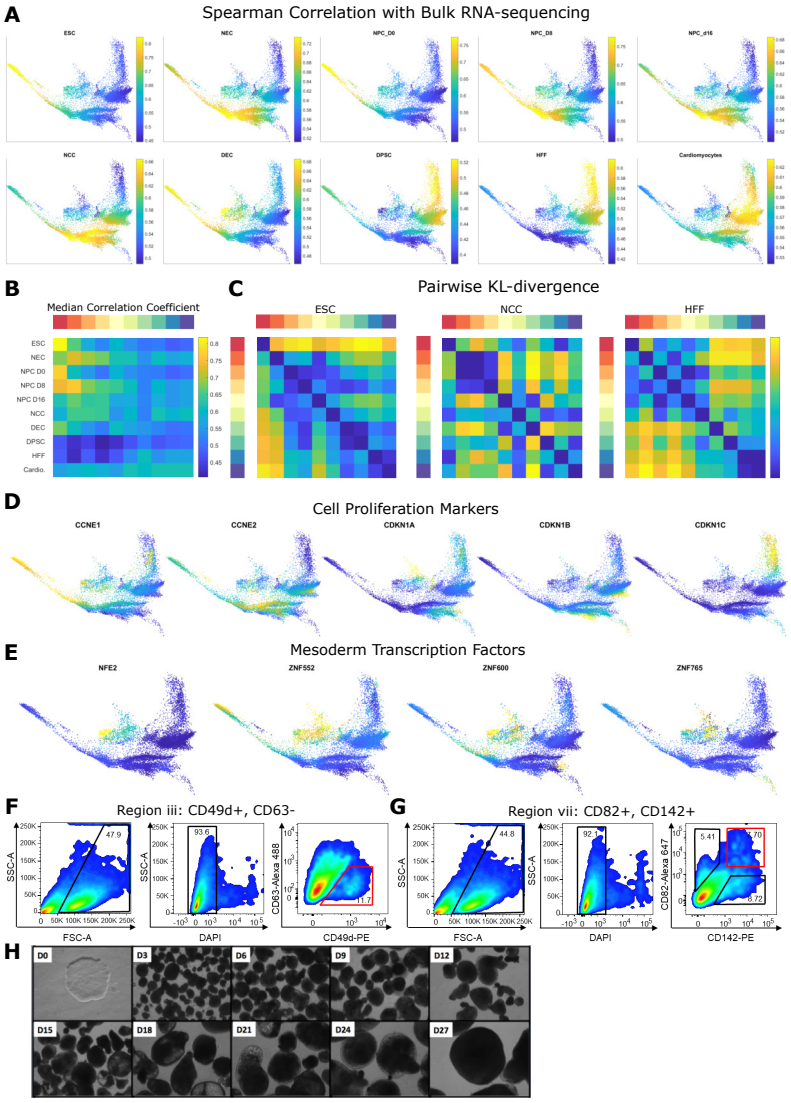
The bulk and scRNA-seq data were compared as follows: For statistical testing, the cells were divided into branches or regions of analysis as shown in Figure 7A (right). For each cell, we calculated the Spearman correlation coefficient between the cell's gene expression levels (after MAGIC [19]) and the bulk gene expression levels. As shown in Figures S7A and S7B, the regions with the highest correlation correspond well with the predicted regions for each cell type.

To determine if the expected region for each cell type in the bulk RNA-seq samples has significantly higher correlation than other regions, we computed a Kullback-Leibler (KL) divergence-based p-value (see Methods). For all the comparisons, the largest resulting p-value was 3.3×10^{-6}, indicating that the correlation coefficient distributions of all regions are statistically different from each other (Figure S7C). The resulting matrices are consistent with the visualizations in Figure S7A. Our comparison to bulk RNA-seq data therefore confirms that the branches within the PHATE visualization are associated with correct cell types and the overall structure in the visualization is correct.

4.4 Experimental Validation of PHATE-Identified Lineages

PHATE visuals show significant structure in datasets including different branches, which can correspond to lineages in developmental data. Therefore, such visualizations can inform biologists about population structures in the data. Additionally, the ability to extract regions of the data such as branches allows us to identify specific gene markers that can be used to experimentally isolate these cell populations for further study. Here, we identify a set of surface markers for the isolation and molecular characterization of cell populations within the EB differentiation process.

We focused on the neural crest and the NCC-derived ectomesenchyme lineages for isolation by deriving surface markers for FACS sorting. We first performed differential expression analysis on the single-cell data by comparing each surface marker expression distribution within the region of interest to the surface marker expression distribution in all other cells (i.e. the "background" expression distribution) using the earth mover's distance (EMD). EMD is a measure



Supplemental Figure S7: Projecting bulk RNA-seq measurements into single-cell resolution. (**A**) PHATE visualization with cells colored by the Spearman correlation coefficients with the bulk RNA-seq data sets. (**B**) Median Spearman correlation coefficient between the cells' TF expression level with the TF expression levels of the bulk RNA-seq datasets. (**C**) Estimated pairwise KL-divergence between the distributions of correlation coefficients in the regions of analysis for selected datasets. All nonzero values are statistically greater than zero (all p-values $< 3.3 \times 10^{-6}$). The resulting matrices are consistent with (**A**). For example, for the ESC bulk sample, the correlation coefficient is very high in region i and relatively low in all the other regions. The KL divergence between the correlation coefficient distributions of region i and all other regions is shown to be high. Similar results were obtained in the other samples and when estimating the Renyi-α divergence for $\alpha = 0.5$. (**D**) PHATE colored by selected proliferation markers. (**E**) PHATE colored by transcription factors that we identified to be associated with the mesoderm that were not known in the literature. (**F**) Scatter plots showing the gating procedure for FACS sorting cell populations of region iii (CD49d and CD63). (**G**) Same as **F** for region vii (CD82 and CD142). (**H**) 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.

of dissimilarity between probability distributions [61]. The resulting scores are shown in Figure 7C. We then selected the highest scoring surface markers that distinctively identify NCC and NCC-derived ectomesenchyme regions. Based on these analyses and the availability of antibodies, CD49D/ITGA4 was chosen for the neural crest while CD142/F3, CD63, and CD82 were chosen for the NCC-derived mesenchyme. Using these surface markers, we FACS-purified cell populations $CD49d^{+}/CD63^{-}$ and $CD82^{+}/CD142^{+}$ and performed bulk RNA-sequencing (Figures S7F & S7G) on these sorted populations.

To verify that we isolated the correct regions of interest, we calculated the Spearman correlation between all PHATE branches and the bulk RNA-seq data from the $CD49d^{+}/CD63^{-}$ sorted cells (Figure 7E). The correlation coefficient was the highest in the neural crest branch (region iii), which corresponds to the highest expression of CD49d. Similar results were obtained for the NCC-derived ectomesenchyme branch in region v (Figure 7F).

Finally, we compared the transcription factor expression levels in the bulk RNA-seq data for the two sorted populations to the mean single-cell RNA-seq data within the corresponding regions (Figure 7G). The correlation coefficients were 0.65 and 0.76 for the neural crest and ectomesenchyme branches, respectively, further validating the correspondence. These analyses highlight the power of PHATE in guiding experimental design.

4.5 Inferring New Genetic Associations with PHATE

To uncover the transcriptional programs underlying hESC differentiation we next identified transcription factors (TFs) that are uniquely expressed in the different regions of the PHATE visualization using the same approach as for the surface markers. TF signatures for the ESC, neural crest, neural progenitor, and mesendoderm branches (regions i-v) were highly specific indicating that differentiation along the respective branches is driven, at least in part, by the dynamic changes in TF networks. In contrast, the NCC-derived ectomesenchymal branches (regions vi-x) exhibited highly overlapping TF signatures consistent with their common neural crest origin. TF profiles of these regions suggest that late-stage differentiation of neural crest derivatives may rely on fine-tuning the levels of multiple TFs and possibly on post-transcriptional mechanisms.

Importantly, while germ layer development in humans has not been explored in great detail, a number of regulators in our list have been implicated in germ layer development in murine models and map to correct differentiation branches of the PHATE visualization, suggesting a conservation of major developmental pathways in human. Among such genes are the core ESC regulators NANOG and POU5F1/ OCT4, the neural progenitor TFs SOX1 and ZIC1, neural crest TFs PAX3 and PAX7 and others. In addition, we have identified a large cohort of novel genes that have not been implicated in germ layer development. For example, the zinc finger proteins ZNF552, ZNF600, and ZNF765 exhibit high and specific expression in the mesendoderm region and thus may play a role in differentiation along this branch (Figure S7E).

The TFs that are associated with such branches can be the subject of further experimentation using perturbations. Thus the ability of PHATE to associate novel genes with lineages can be a powerful tool for biological inquiry.

4.6 PHATE Unveils Large Scale Gene Expression Patterns

In addition to the characterization of specific lineages, the comprehensive picture provided by PHATE can also be used to learn larger scale patterns of gene expression. To obtain a comprehensive view of the molecular mechanisms that regulate germ layer development in the hESC differentiation model we analyzed genes dynamically expressed along different branches of the PHATE visualization, focusing on proliferation signatures, transcription factors, and chromatin modifiers.

Changes in the epigenetic state of a cell are thought to play a key role in differentiation. Indeed, our analyses identified a set of chromatin modifiers specifically upregulated in undifferentiated ESCs. Overall, however, chromatin modifier signatures exhibited gradual transitions along differentiation branches. For example, the undifferentiated hESC and neuroectoderm progenitor branches (regions i and ii, respectively) share a number of chromatin modifiers, despite having sharply different TF profiles. Likewise, common chromatin modification signatures were observed for the neural and neural crest progenitor branches (regions iii and iv, respectively).

It has also been observed that proliferation inversely correlates with differentiation state, being the highest in stem/progenitor cell compartments and the lowest in terminally differentiated cells. ESCs have a very short G1 phase and are characterized by a rapid cell cycle dividing once every twelve hours. hESCs undergoing differentiation lengthen G1 phase and increase doubling time which reaches 24-48h for most somatic cells in culture [70]. To visualize proliferation rate along the branches of the PHATE visualization, we examined the expression of Cyclin E family members CCNE1/2 which are expected to be up-regulated in rapidly proliferating cells as well as expression of CDK inhibitors CDKN1A/B/C and CDKN2A/B/C which negatively regulate cell cycle progression. Indeed, CCNE1/2 expression was highest in the branches containing hESCs and neuroectoderm precursors (regions i and ii, respectively) and in the early neural and neural crest precursors. Expression of these markers declined sharply at the end points of all branches. Conversely, expression of CDK inhibitors was highest at the branch end points, indicating the presence of differentiated cells (Figure S7D).

Taken together, our analyses demonstrate the power of the PHATE visualization to generate comprehensive and dynamic molecular profiles in complex differentiation systems that has the potential to greatly accelerate the pace of biological discoveries.

5 Conclusion

With large amounts of high dimensional, high-throughput biological data being generated in many types of biological systems, there is a great need for interpretable visualizations that can represent structures in data without strong prior assumptions. However, most existing methods are highly deficient at retaining structures of interest in biology such as clusters, trajectories or progressions of various dimensionality, hybrids of the two as well as local and global nonlinear relationships in data. Further, they have trouble contending with the size of modern datasets and

to the noise inherent to biological datasets. PHATE provides a unique solution to these problems by creating a diffusion-based informational geometry from the data, and preserving a divergence metric between datapoints that is sensitive to near and far manifold-intrinsic distances in the dataspace. The information geometry created in PHATE is based on data diffusion dynamics, which are able to quantify distances between points as differences in probabilities of t-step random walks. Such probabilities are known to be robust to noise, but due to the harmonics of diffusion, they are sensitive to many scales of differences (both large and small) in signals, thus revealing intricate local structure as well as global structure in a denoised way.

We applied PHATE to a wide variety of datasets including single-cell data such as single-cell RNA sequencing, and CyTOF, in addition to Gut Microbiome and SNP data where the units are patients and not cells. We also tested PHATE on network data such as Hi-C, Facebook, and Wikipedia networks. In each case we see that PHATE is able to reveal structures of visual interest to humans that other methods entirely miss. We have also implemented PHATE in such a way that it processes millions of datapoints in a matter of hours. Thus currently PHATE is the only method that can efficiently handle the datasets that are now being produced using single-cell RNA sequencing technologies, as we demonstrated using the 1.3 million cell dataset released by 10X genomics.

We demonstrated the application of PHATE to biological data exploration on our newly generated EB differentiation system. Here, we found that PHATE successfully resolves cell heterogeneity and correctly maps differentiation trajectories in complex systems based on scRNA-seq data alone, without any additional assumptions on the data. Furthermore, gene expression signatures associated with specific differentiation branches can be extracted and used to learn key biological features of specific cell populations. The insights obtained with PHATE could be particularly valuable for identifying novel markers for purification of transient differentiation stages for further molecular and functional analyses. Thus, PHATE will be particularly useful for dissecting the molecular mechanisms regulating stem and progenitor cell compartments in different tissues that are not easily accessible via conventional approaches.

We expect numerous biological, but also non-biological, data types to benefit from PHATE, including applications in high-throughput genomics, phenotyping, and many other fields. We believe that PHATE will revolutionize biomedical data exploration by offering a new way of visualizing and extracting information from high-dimensional data.

Branch	Clusters	Cell Types
A	27, 2, 8, 22, 10, 12	ESC/neuroectoderm
B	19, 4, 21	Neural crest cells
C	30, 18, 16, 13, 15, 23	Neural progenitor cells
D	7, 11, 25, 1	Mesendoderm
E	24, 28, 20, 14, 6	Ectomesenchymal progenitors w/ heart component
F	26, 3	Ectomesenchymal progenitors w/ bone component
G	9, 29	Ectomesenchymal neural crest derivatives

Supplemental Table S2: Cluster numbers from Figure 7A (middle left) that comprise the branches in Figure S3E.

References

[1] 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.

[2] 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.

[3] 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.

[4] X. Qiu, Q. Mao, Y. Tang, L. Wang, R. Chawla, H. A. Pliner, and C. Trapnell, "Reversed graph embedding resolves complex single-cell trajectories.," *Nature Methods*, 2017.

[5] 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.

[6] J. B. Tenenbaum, V. De Silva, and J. C. Langford, "A global geometric framework for nonlinear dimensionality reduction," *science*, vol. 290, no. 5500, pp. 2319–2323, 2000.

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

[8] R. Talmon and R. R. Coifman, "Empirical intrinsic geometry for nonlinear modeling and time series filtering," *Proceedings of the National Academy of Sciences*, vol. 110, no. 31, pp. 12535–12540, 2013.

[9] R. Talmon and R. R. Coifman, "Intrinsic modeling of stochastic dynamical systems using empirical geometry," *Applied and Computational Harmonic Analysis*, vol. 39, no. 1, pp. 138 – 160, 2015.

[10] M. Salicrú and A. A. Pons, "Sobre ciertas propiedades de la m-divergencia en análisis de datos," *Qüestiió: quaderns d'estadística i investigació operativa*, vol. 9, no. 4, pp. 251–256, 1985.

[11] M. Salicrú, A. Sanchez, J. Conde, and P. Sanchez, "Entropy measures associated with K and M divergences," *Soochow Journal of Mathematics*, vol. 21, no. 3, pp. 291–298, 1995.

[12] E. Hellinger, "Neue begründung der theorie quadratischer formen von unendlichvielen veränderlichen.," *Journal für die reine und angewandte Mathematik*, vol. 136, pp. 210–271, 1909.

[13] K. Shekhar, S. W. Lapan, I. E. Whitney, N. M. Tran, E. Z. Macosko, M. Kowalczyk, X. Adiconis, J. Z. Levin, J. Nemesh, M. Goldman, *et al.*, "Comprehensive classification of retinal bipolar neurons by single-cell transcriptomics," *Cell*, vol. 166, no. 5, pp. 1308–1323, 2016.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] 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.

[20] X. Cheng, M. Rachh, and S. Steinerberger, "On the diffusion geometry of graph laplacians and applications," *arXiv preprint arXiv:1611.03033*, 2016.

[21] 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.

[22] E. Levina and P. J. Bickel, "Maximum likelihood estimation of intrinsic dimension," in *Advances in neural information processing systems*, pp. 777–784, 2005.

[23] 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.

[24] Y. Tsukamoto and N. Omi, "Classification of mouse retinal bipolar cells: type-specific connectivity with special reference to rod-driven aii amacrine pathways," *Frontiers in neuroanatomy*, vol. 11, p. 92, 2017.

[25] 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.

[26] J. H. Levine, E. F. Simonds, S. C. Bendall, K. L. Davis, D. A. El-ad, M. D. Tadmor, O. Litvin, H. G. Fienberg, A. Jager, E. R. Zunder, *et al.*, "Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis," *Cell*, vol. 162, no. 1, pp. 184–197, 2015.

[27] 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.

[28] S. T. Roweis and L. K. Saul, "Nonlinear dimensionality reduction by locally linear embedding," *science*, vol. 290, no. 5500, pp. 2323–2326, 2000.

[29] E. M. Darrow, M. H. Huntley, O. Dudchenko, E. K. Stamenova, N. C. Durand, Z. Sun, S.-C. Huang, A. L. Sanborn, I. Machol, M. Shamim, A. P. Seberg, E. S. Lander, B. P. Chadwick, and E. Lieberman Aiden, "Deletion of dxz4 on the human inactive x chromosome alters higher-order genome architecture," *Proceedings of the National Academy of Sciences*, p. 201609643, 2016.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

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

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] D. S. Kim, S. E. Ross, J. M. Trimarchi, J. Aach, M. E. Greenberg, and C. L. Cepko, "Identification of molecular markers of bipolar cells in the murine retina," *Journal of Comparative Neurology*, vol. 507, no. 5, pp. 1795–1810, 2008.

[41] "Our 1.3 million single cell dataset is ready to download," Feb. 2017.

[42] C. Klymko, D. Gleich, and T. G. Kolda, "Using triangles to improve community detection in directed networks," *arXiv preprint arXiv:1404.5874*, 2014.

[43] G. R. Martin and M. J. Evans, "Differentiation of clonal lines of teratocarcinoma cells: formation of embryoid bodies in vitro," *Proceedings of the National Academy of Sciences*, vol. 72, no. 4, pp. 1441–1445, 1975.

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *Computer Vision, 1998. IEEE Sixth International Conference on*, pp. 59–66, IEEE, 1998.

[62] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and computing*, vol. 17, no. 4, pp. 395–416, 2007.

[63] A. Y. Ng, M. I. Jordan, and Y. Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in neural information processing systems*, pp. 849–856, 2002.

[64] L.-F. Chu, N. Leng, J. Zhang, Z. Hou, D. Mamott, D. T. Vereide, J. Choi, C. Kendziorski, R. Stewart, and J. A. Thomson, "Single-cell rna-seq reveals novel regulators of human embryonic stem cell differentiation to definitive endoderm," *Genome Biology*, vol. 17, no. 1, p. 173, 2016.

[65] P. Freire-Pritchett, S. Schoenfelder, C. Várnai, S. W. Wingett, J. Cairns, A. J. Collier, R. García-Vílchez, M. Furlan-Magaril, C. S. Osborne, P. Fraser, *et al.*, "Global reorganisation of cis-regulatory units upon lineage commitment of human embryonic stem cells," *eLife*, vol. 6, p. e21926, 2017.

[66] Y. Qiao, X. Wang, R. Wang, Y. Li, F. Yu, X. Yang, L. Song, G. Xu, Y. E. Chin, and N. Jing, "Af9 promotes hesc neural differentiation through recruiting tet2 to neurodevelopmental gene loci for methylcytosine hydroxylation," *Cell Discovery*, vol. 1, p. 15017, 2015.

[67] A. Rada-Iglesias, R. Bajpai, S. Prescott, S. A. Brugmann, T. Swigut, and J. Wysocka, "Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest," *Cell stem cell*, vol. 11, no. 5, pp. 633–648, 2012.

[68] N. Urraca, R. Memon, I. El-Iyachi, S. Goorha, C. Valdez, Q. T. Tran, R. Scroggs, G. A. Miranda-Carboni, M. Donaldson, D. Bridges, *et al.*, "Characterization of neurons from immortalized dental pulp stem cells for the study of neurogenetic disorders," *Stem cell research*, vol. 15, no. 3, pp. 722–730, 2015.

[69] Q. Liu, C. Jiang, J. Xu, M.-T. Zhao, K. Van Bortle, X. Cheng, G. Wang, H. Y. Chang, J. C. Wu, and M. P. Snyder, "Genome-wide temporal profiling of transcriptome and open chromatin of early cardiomyocyte differentiation derived from hipscs and hescsnovelty and significance," *Circulation Research*, vol. 121, no. 4, pp. 376–391, 2017.

[70] K. Kapinas, R. Grandy, P. Ghule, R. Medina, K. Becker, A. Pardee, S. K. Zaidi, J. Lian, J. Stein, A. van Wijnen, *et al.*, "The abbreviated pluripotent cell cycle," *Journal of cellular physiology*, vol. 228, no. 1, pp. 9–20, 2013.

[71] 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.

[72] 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.

[73] 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.

[74] 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.

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

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

[77] 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.

[78] D. Kaplan, "Knee Point - File Exchange - MATLAB Central," 2012.

[79] S. Eguchi *et al.*, "A differential geometric approach to statistical inference on the basis of contrast functionals," *Hiroshima mathematical journal*, vol. 15, no. 2, pp. 341–391, 1985.

[80] I. Csiszár, "Eine informationstheoretische ungleichung und ihre anwendung auf beweis der ergodizitaet von markoffschen ketten," *Magyer Tud. Akad. Mat. Kutato Int. Koezl.*, vol. 8, pp. 85–108, 1964.

[81] S. M. Ali and S. D. Silvey, "A general class of coefficients of divergence of one distribution from another," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 131–142, 1966.

[82] L. M. Bregman, "The relaxation method of finding the common point of convex sets and its application to the solution of problems in convex programming," *USSR computational mathematics and mathematical physics*, vol. 7, no. 3, pp. 200–217, 1967.

[83] S. Kullback and R. A. Leibler, "On information and sufficiency," *The annals of mathematical statistics*, vol. 22, no. 1, pp. 79–86, 1951.

[84] S. Amari, *Information geometry and its applications*. Springer, 2016.

[85] S.-i. Amari and H. Nagaoka, *Methods of information geometry*, vol. 191 of *Translations of Mathematical Monographs*. American Mathematical Society, Providence, RI; Oxford University Press, Oxford, 2000. Translated from the 1993 Japanese original by Daishi Harada.

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

[87] G. Wolf, A. Rotbart, G. David, and A. Averbuch, "Coarse-grained localized diffusion," *Applied and Computational Harmonic Analysis*, vol. 33, no. 3, pp. 388 – 400, 2012.

[88] 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.

[89] K. R. Moon and A. O. Hero, "Ensemble estimation of multivariate f-divergence," in *Information Theory (ISIT), 2014 IEEE International Symposium on*, pp. 356–360, IEEE, 2014.

[90] K. Moon and A. Hero, "Multivariate f-divergence estimation with confidence," in *Advances in Neural Information Processing Systems*, pp. 2420–2428, 2014.

[91] K. Treleaven and E. Frazzoli, "An explicit formulation of the earth movers distance with continuous road map distances," *arXiv preprint arXiv:1309.7098*, 2013.

[92] K. Q. Weinberger, F. Sha, and L. K. Saul, "Learning a kernel matrix for nonlinear dimensionality reduction," in *Proceedings of the twenty-first international conference on Machine learning*, p. 106, ACM, 2004.

[93] A. L. Haber, M. Biton, N. Rogel, R. H. Herbst, K. Shekhar, C. Smillie, G. Burgin, T. M. Delorey, M. R. Howitt, Y. Katz, *et al.*, "A single-cell survey of the small intestinal epithelium," *Nature*, vol. 551, no. 7680, p. 333, 2017.

[94] S. A. Nene, S. K. Nayar, H. Murase, *et al.*, "Columbia object image library (coil-20)," 1996.

[95] Y. LeCun, "The mnist database of handwritten digits," *http://yann. lecun. com/exdb/mnist/*, 1998.

[96] 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.

[97] B. Anchang, T. D. Hart, S. C. Bendall, P. Qiu, Z. Bjornson, M. Linderman, G. P. Nolan, and S. K. Plevritis, "Visualization and cellular hierarchy inference of single-cell data using spade," *Nature protocols*, vol. 11, no. 7, pp. 1264–1280, 2016.

[98] T. K. S. Moon and C. Wynn, *Mathematical methods and algorithms for signal processing*. Prentice Hall, 2000.

[99] 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.

[100] M. Wattenberg, F. Viégas, and I. Johnson, "How to use t-sne effectively," *Distill*, 2016.

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

[102] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische mathematik*, vol. 1, no. 1, pp. 269–271, 1959.

[103] R. W. Floyd, "Algorithm 97: shortest path," *Communications of the ACM*, vol. 5, no. 6, p. 345, 1962.

[104] L. Van Der Maaten, E. Postma, and J. Van den Herik, "Dimensionality reduction: A comparative review," *Journal of Machine Learning Research*, vol. 10, pp. 66–71, 2009.

[105] M. Balasubramanian and E. L. Schwartz, "The isomap algorithm and topological stability," *Science*, vol. 295, no. 5552, pp. 7–7, 2002.

[106] J. A. Lee and M. Verleysen, "Nonlinear dimensionality reduction of data manifolds with essential loops," *Neurocomputing*, vol. 67, pp. 29–53, 2005.

[107] I. S. Lim, P. de Heras Ciechomski, S. Sarni, and D. Thalmann, "Planar arrangement of high-dimensional biomedical data sets by isomap coordinates," in *Computer-Based Medical Systems, 2003. Proceedings. 16th IEEE Symposium*, pp. 50–55, IEEE, 2003.

[108] A. A. Hagberg, D. A. Schult, and P. J. Swart, "Exploring network structure, dynamics, and function using NetworkX," in *Proceedings of the 7th Python in Science Conference (SciPy 2008)* (G. Varoquaux, T. Vaught, and J. Millman, eds.), (Pasadena, CA USA), pp. 11 – 15, 2008.

[109] T. M. Fruchterman and E. M. Reingold, "Graph drawing by force-directed placement," *Software: Practice and experience*, vol. 21, no. 11, pp. 1129–1164, 1991.

[110] Q. Mao, L. Wang, S. Goodison, and Y. Sun, "Dimensionality reduction via graph structure learning," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 765–774, ACM, 2015.

[111] Q. Mao, L. Wang, I. Tsang, and Y. Sun, "Principal graph and structure learning based on reversed graph embedding," *IEEE transactions on pattern analysis and machine intelligence*, 2016.

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

[113] 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.

[114] Y. Hart, H. Sheftel, J. Hausser, P. Szekely, N. B. Ben-Moshe, Y. Korem, A. Tendler, A. E. Mayo, and U. Alon, "Inferring biological tasks using pareto analysis of high-dimensional data," *Nature methods*, vol. 12, no. 3, pp. 233–235, 2015.

[115] O. Shoval, H. Sheftel, G. Shinar, Y. Hart, O. Ramote, A. Mayo, E. Dekel, K. Kavanagh, and U. Alon, "Evolutionary trade-offs, pareto optimality, and the geometry of phenotype space," *Science*, vol. 336, no. 6085, pp. 1157–1160, 2012.

[116] E. Lieberman-Aiden, N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander, and J. Dekker, "Comprehensive mapping of long-range interactions reveals folding principles of the human genome," *Science*, vol. 326, no. 5950, pp. 289–293, 2009.

[117] E. P. Consortium *et al.*, "An integrated encyclopedia of dna elements in the human genome," *Nature*, vol. 489, no. 7414, pp. 57–74, 2012.

[118] J. Leskovec and J. J. Mcauley, "Learning to discover social circles in ego networks," in *Advances in neural information processing systems*, pp. 539–547, 2012.

[119] 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.

[120] 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.

6 Methods

Here we present an expanded explanation of our computational methods, experimental methods and data processing steps.

6.1 Computational Methods

6.1.1 Manifold, Diffusion, and Information Geometry Data Models

To establish an abstract geometric model for high-dimensional data visualized by PHATE, we consider two properties that we typically observed in high throughput data (biomedical and otherwise). First, transitions between datapoints tend to be incremental and gradual. There may be many such patches of incremental change but nevertheless these gradual transitions are usually prevalent. Secondly, there are a limited number of intrinsic directions (or pathways) along which datapoints progress. Therefore, the dynamics captured by collected data are inherently more like a set of rivers, rather than a cloud (expanding outwards in all directions).

Data with such properties can thus be modeled geometrically by a collection of smoothly varying data patches defined by local neighborhoods. This collection essentially fits the manifold learning paradigm, which relies on a mathematical manifold model for the geometry of a progression track, together with analysis tools for characterizing it. Furthermore, data manifolds often have a low intrinsic dimension, even if curvature and noise forces them to span a high dimensional ambient volume in the collected feature space. Finally, progression tracks form trajectories, with a limited number of "branching points", where progression splits into several directions. Therefore, in this case the underlying data geometry can implicitly be regarded as a collection of intrinsically low-dimensional manifolds (i.e., curves, surfaces) that cross each other in branching points.

It has been shown in several works (e.g., [71, 72]) that manifold geometries are closely related to heat diffusion, which is modeled by the heat equation – a differential equation defined in terms of the Laplace-Beltrami operators. Indeed, meta-stable 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 [7, 73, 74]. The embedding obtained by PHATE extends these results by considering this diffusion goemetry as a statistical manifold of diffusion distributions and using tools of information geometry (namely, α-representations) to capture its metric stucture and embed it in visualizable (i.e., two or three) dimensions. Further, as we discuss in the following sections, the information distance metric we use also relates to Boltzman energy potentials of the diffusion process, and therefore it combines together both the dyamical systems and information geometry aspects of data-driven diffusion geometries. In particular, for the case of transition structures, this approach enables the consideration of underlying data geometry consisting of multiple low-dimensional manifolds (such as trajectory curves) that cross each other, while alleviating boundary-condition instabilities to maintain low dimensionality of the embedded space that is better-suited for visualization. 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.

6.1.2 The Diffusion Operator

Here we discuss the construction of the diffusion operator. PHATE is based on constructing a diffusion geometry to learn and represent the shape of the data [7, 73, 74]. 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 local similarities or affinities, typically via some kernel function (e.g. a Gaussian kernel).

Let $\mathcal{X} \subset \mathbb{R}^d$ be a dataset with N points sampled i.i.d. from a probability distribution $p : \mathbb{R}^d \to [0, \infty)$ (with $\int p(x)dx = 1$) that is essentially supported on a low dimensional manifold $\mathcal{M}^m \subseteq \mathbb{R}^d$, where m is the dimension of $\mathcal{M}$ and $m \ll d$. The classic diffusion geometry proposed in [7] 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

$$\nu_\varepsilon(x) = \|k_\varepsilon(x, \cdot)\|_1 = \sum_{z \in \mathcal{X}} k_\varepsilon(x, z) \tag{1}$$

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 $\Pr[x \to y] = [P_\varepsilon]_{(x,y)}$.

The α-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 distance, along with an α-decaying kernel that controls the rate of decay of the kernel.

The original heuristic proposed in [7] 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, resulting in a high affinity with many points that are far away. 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 [75]. See Figure S2D 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 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 too sparsely connected, making it difficult to learn long range connections. Thus we recommend that α be fixed at a large number (e.g. $\alpha \geq 10$) and then k can be chosen sufficiently large to ensure that points are locally connected. 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.

In addition to progression or trajectory structures, the recommendations provided in this section work well for visualizing data that naturally separate into distinct clusters. In particular, the α-decay kernel ensures that relationships are preserved between distinct clusters that are relatively close to each other.

6.1.3 Powering the Diffusion Operator

Here we discuss diffusion, i.e., raising the diffusion operator to its t-th power as shown 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 described 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 [74] 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

$$D^t(x,y) \triangleq \|p_x^t - p_y^t\|_{\ell_2(1/p^\infty)} = \left(\sum_{z \in \mathcal{X}} \frac{(p_x^t(z) - p_y^t(z))^2}{p^\infty(z)} \right)^{1/2} , \tag{5}$$

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 $D^t(x,y) = \|\Phi^t(x) - \Phi^t(y)\|_2$.

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 formulate a new algorithm for choosing the timescale t. Our algorithm quantifies the information in the powered diffusion operator with various values of t. This is accomplished by computing the spectral or *Von Neumann Entropy* (VNE) [76, 77] 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 VNE 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 around 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 and for computational ease, 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. The VNE of this diffusion affinity is used to quantify the amount of variability. 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 (and equivalently of P^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 plot of Figure S2E. Choosing t from among the values where $H(t)$ is decreasing rapidly generally results in noisy visualizations and

embeddings (second plot in Figure S2E). Very large values of t result in a visualization where some of the branches or trajectories are combined together and some of the signal is lost (fourth plot in Figure S2E). 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 around the "knee" in the plot of $H(t)$ (third plot in Figure S2E and the PHATE visualizations in Figure 1). 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 visualization is fairly robust to the choice of t in this range, as demonstrated in Figure S2F. In the code, we include an automatic method for selecting t based on a knee point detection algorithm that finds the knee by fitting two lines to the VNE curve [78].

6.1.4 Potential Distances

In order to provide mathematical context to the potential distance used in PHATE, we relate it here to the heat propagation dynamics that govern the diffusion geometry and the diffusion process we use to build it. This heat diffusion process can be analyzed by considering two possible scenarios for the origin of the dataset $\mathcal{X}$ and its distribution p, as described in [73, 74]. 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 [73, 74] 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 [73, 74], 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* $\mathfrak{V}^t(x,y) = \left(\sum_{z\in\mathcal{X}} \log^2\left(\frac{1+\left\langle \Phi^{t/2}(x), \Phi^{t/2}(z)\right\rangle}{1+\left\langle \Phi^{t/2}(y), \Phi^{t/2}(z)\right\rangle}\right)\right)^{1/2}$

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

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

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

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

$$\begin{aligned}\left(U^t_{\varepsilon,x}(x_j) - U^t_{\varepsilon,y}(x_j)\right)^2 &= \left[\log(1 + \left\langle \Phi^{t/2}_\varepsilon(x), \Phi^{t/2}_\varepsilon(x_j)\right\rangle)\right. \\ &- \left.\log(1 + \left\langle \Phi^{t/2}_\varepsilon(y), \Phi^{t/2}_\varepsilon(x_j)\right\rangle)\right]^2 \\ &= \log^2\left(\frac{1 + \left\langle \Phi^{t/2}_\varepsilon(x), \Phi^{t/2}_\varepsilon(x_j)\right\rangle}{1 + \left\langle \Phi^{t/2}_\varepsilon(y), \Phi^{t/2}_\varepsilon(x_j)\right\rangle}\right),\end{aligned}$$

which yields the result in the proposition. □

6.1.5 Diffusion-based Informational Distances

The potential distance can be generalized to a family of diffusion-based distances rooted in an information geometry interpretation of the diffusion geometry. Here we expand the discussion of this aspect to establish the relation between the diffusion distance (Eq. 5) and the potential distance (Def. 1) as two extremes in this family of distances. Indeed, these two distance metrics both aim to quantify the difference between diffusion distributions p^t_x and p^t_y that represent data points $x, y \in \mathcal{X}$. Given an intermediate data point $z \in \mathcal{X}$, and a meta-parameter $-1 \leq \gamma \leq 1$, let

$$\Delta^{(\gamma)}_{(x,y)}(z) = -\int_{p^t_x(z)}^{p^t_y(z)} u^{-\frac{\gamma+1}{2}} du = \begin{cases} p^t_x(z) - p^t_y(z) & \gamma = -1 \\ \log p^t_x(z) - \log p^t_y(z) & \gamma = +1 \\ \frac{2}{1-\gamma}\left[(p^t_x(z))^{\frac{1-\gamma}{2}} - (p^t_y(z))^{\frac{1-\gamma}{2}}\right] & \text{otherwise} \end{cases} \tag{8}$$

quantify the difference between the transition probabilities $p^t_x(z), p^t_y(z)$ from x, y to z. Then, diffusion distances and potential distances are given by L^2 norms of $\Delta^{(-1)}_{(x,y)}$ and $\Delta^{(+1)}_{(x,y)}$ correspondingly (albeit with a different measure over z, since diffusion distances are defined over $L^2(\frac{1}{p^\infty})$ in Eq. 5). Therefore, these distances can be regarded as two extremes of a general family of distances over the diffusion geometry. Moreover, as we show in Prop 2, diffusion dissimilarities of the form $\|\Delta^{(\gamma)}_{(x,y)}\|_2$ combine both this diffusion-geometry notion of a data-driven distance metric, and an information-theory notion of divergence between diffusion distributions, which is provided in Def. 2 for completeness.

Definition 2 (Divergence [Information Theory])**.** *Let S be a space of probability distributions with common support. A divergence on S is a function $D(\cdot||\cdot) : S \times S \to \mathbb{R}$ s.t. (1) $D(p||q) \geq 0$ for all $p, q \in S$, and (2) $D(p||q) = 0$ if and only if $p = q$* [79]*. Some specific classes of divergences include:*

(i) M-divergence [10, 11]: *Let p and q be probability mass functions and let g be a differentiable function with continuous second derivative. Then the M-divergence between p and q is $M_g(p, q) = \sum_i (g(p_i) - g(q_i))^2$.*

(ii) f-divergence [80, 81]: *Let f be a convex function s.t. $f(1) = 0$ and let p and q be probability mass functions. The f-divergence between p and q is $D_f(p||q) = \sum_i q_i f\left(\frac{p_i}{q_i}\right)$.*

(iii) Bregman divergence [82]: *Let Ω be a closed convex set and let $F : \Omega \to \mathbb{R}$ be a strictly convex and continuously differentiable function. The Bregman divergence between the points $p, q \in \Omega$ is $D_F(p, q) = F(p) - F(q) - \langle \nabla F(q), p - q \rangle$, where $\nabla F(q)$ is the gradient of F evaluated at q.*

Unlike (formal) distance metrics, a divergence is not required to be symmetric nor satisfy the triangle inequality. Examples of f-divergences include the Kullback-Leibler divergence [83] and the Hellinger distance [12]. An example of a Bregman divergence is the squared Euclidean distance where $F(x) = x^2$. Note that these different types of divergences are not mutually exclusive as is evident in the following proposition.

Proposition 2. *The squared norm $\|\Delta^{(\gamma)}_{(x,y)}\|^2$ forms an M-divergence between the diffusion probability distributions p^t_x and p^t_y for all $\gamma \in [-1, 1]$. Furthermore, $\|\Delta^{(\gamma)}_{(x,y)}\|^2$ forms an* f*-divergence for $\gamma = 0$ and a Bregman divergence for $\gamma = -1$.*

Proof. For $\gamma \in (-1, 1)$, it follows that $\|\Delta^{(\gamma)}_{(x,y)}\|^2$ is an M-divergence from Definition 2 with $g(x) = \frac{2}{1-\gamma} x^{\frac{1-\gamma}{2}}$. Similarly, $g(x) = x$ and $g(x) = \log x$ yield this result for $\gamma = -1, 1$, respectively. For $\gamma = 0$, we obtain $\|\Delta^{(0)}_{(x,y)}\|^2 = 4 \sum_z \left[\sqrt{p^t_x(z)} - \sqrt{p^t_y(z)}\right]^2$, which is indeed an f-divergence as it is proportional to the Hellinger divergence. Finally, since $\gamma = -1$ yields a squared Euclidean distance between the distributions it is indeed a Bregman divergence. □

The family of distances (or divergences) formed by $\|\Delta^{(\gamma)}_{(x,y)}\|_2$ is also directly related to α-representations used in information geometries [84] when defining statistical manifolds over probability distributions. Indeed, the α-representation of a distribution p is defined as

$$\ell^{(\alpha)}(p) = \begin{cases} \frac{2}{1-\alpha} p^{\frac{1-\alpha}{2}} & \alpha \neq 1 \\ \log p & \alpha = 1, \end{cases}$$

where $\ell^{(-1)}$ and $\ell^{(+1)}$ give rise to the popular mixed family (m-family) and exponential family (e-family) in information geometry [84, 85], correspondingly. We also note at this point that the third popular α-family in information geometry is the 0-family, which gives a Fisher geometry as its Riemannian metric is given by Fisher information [85]. Interestingly, this family corresponds to setting $\gamma = 0$ in our case, which yields a distance $\|\Delta^{(0)}_{(x,y)}\|_2 \propto \|\sqrt{p^t_x} - \sqrt{p^t_y}\|_2$ that is proportional to the Hellinger distance between diffusion distributions.

Since, as we discussed here, $\Delta^{(\gamma)}_{(x,y)}$ encodes differences between information geometry $\ell^{(\gamma)}$ representations of diffusion distributions, we refer to the distances $\|\Delta^{(\gamma)}_{(x,y)}\|_2$ (from Prop. 2) as diffusion-based informational distances. In general, this family of informational distances

creates an exciting connection between diffusion geometries and information geometries for exploring emergent structures in data exploration. In particular, in PHATE we focus on the potential distance as an e-family distance [84, 85] that combines both the Boltzman distribution law approach and the information geometry approach towards capturing a stable metric structure of the diffusion geometry, and use it for the purpose of visualizing progression by embedding this metric structure in low dimensions.

6.1.6 Diffusion Potential Embedding via MDS

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). Classical MDS (CMDS) [86] takes a distance matrix as input and embeds the data into a lower-dimensional space as follows. The squared potential distance matrix is double centered:

$$B = -\frac{1}{2} J \mathfrak{V}^{t(2)} J, \tag{9}$$

where $\mathfrak{V}^{t(2)}$ is the squared potential distance matrix (i.e. each entry is squared) and $J = I - \frac{1}{N}\mathbf{1}\mathbf{1}^T$ with $\mathbf{1}$ a vector of ones with length N. The CMDS coordinates are then obtained by an eigendecomposition of the matrix B. This is equivalent to minimizing the following "strain" function:

$$\texttt{Strain}(\hat{x}_1, \ldots, \hat{x}_N) = \sqrt{\sum_{i,j} \left(B_{ij} - \langle \hat{x}_i, \hat{x}_j \rangle\right)^2 \Big/ \sum_{i,j} B_{ij}^2}, \tag{10}$$

over embedded m-dimensional coordinates $\hat{x}_i \in \mathbb{R}^m$ of data points in $\mathcal{X}$. We apply CMDS to the potential distances of the data to obtain an initial configuration of the data in low dimension m.

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. Metric MDS relaxes this assumption by only requiring the input distances to be a distance metric. Metric MDS then embeds the data into lower dimensions by minimizing the following "stress" function:

$$\texttt{Stress}(\hat{x}_1, \ldots, \hat{x}_N) = \sqrt{\sum_{i,j} \left(\mathfrak{V}^t_{(x_i,x_j)} - \|\hat{x}_i - \hat{x}_j\|\right)^2 \Big/ \sum_{i,j} \left(\mathfrak{V}^t_{x_i,x_j}\right)^2}. \tag{11}$$

over embedded m-dimensional coordinates $\hat{x}_i \in \mathbb{R}^m$ of data points in $\mathcal{X}$.

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 m is too small to represent the data without distortion. Thus, by choosing the number

of MDS dimensions to be $m = 2$ (or $m = 3$) for visualization purposes, we may trade off distortion in exchange for readily visualizable coordinates. However, 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 visually distinguishable. By using 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. When analyzing the PHATE coordinates (e.g. for clustering or branch detection), we use metric MDS with m chosen to explain most of the variance in the data as determined by the eigenvalues of the diffusion operator (as is done for von Neumann entropy). In this case, minimal distortion is introduced into the analysis.

In some cases, it may be advantageous to relax our assumptions further on the input distances. In this case, non-metric MDS may be used. In contrast with metric MDS, non-metric MDS does not require the input distances to be an actual distance or metric. Non-metric MDS minimizes the differences between a monotonic transformation of the input dissimilarities and the distances in the embedded space. Mathematically, non-metric MDS minimizes the following stress function:

$$\texttt{Stress}'(\hat{x}_1, \ldots, \hat{x}_N) = \sqrt{\sum_{i,j} \left(\mathfrak{V}^t_{(x_i,x_j)} - f(\|\hat{x}_i - \hat{x}_j\|) \right)^2 \Big/ \sum_{i,j} \left(\mathfrak{V}^t_{x_i,x_j} \right)^2}, \qquad (12)$$

where f is a monotonic transformation of the distances between points in the embedded space.

In our experience, the resulting visualizations from metric MDS and non-metric MDS are nearly identical for most datasets. Furthermore, metric MDS is computationally faster than non-metric MDS. Thus, we recommend metric MDS for most problems.

6.1.7 Robustness Analysis of PHATE

Here we show that the PHATE embedding is robust to subsampling and the choice of t.

Robustness to the scale parameter t Here, we show that the PHATE embedding is quite robust to the choice of t. Figure S2F shows the PHATE embedding on the iPSC mass cytometry dataset from [14] with varying scale parameter t. Figure S2F shows that the embeddings for $50 \leq t \leq 200$ are nearly identical. Thus, PHATE is very robust to the scale parameter t. Similar results can be obtained on other datasets.

Robustness to subsampling We demonstrate that the PHATE algorithm is robust to subsampling of the data by running PHATE on the iPSC mass cytometry dataset from [14] with varying subsample sizes N. Figure S2G shows the PHATE embedding for $N = 1000, 2500, 5000, 10000$. Note that the primary branches or trajectories that are visible when $N = 50000$ (Figure S3C) are still visible for all subsamples. Thus, PHATE is robust to the subsampling size. Similar results can be obtained on other datasets.

6.1.8 Scalability of PHATE

Algorithm 2: Scalable PHATE algorithm

Input: Data matrix X, neighborhood size k, locality scale α, desired embedding dimension m (usually 2 or 3 for visualization), number of landmarks M

Output: The PHATE embedding Y_{points}

1: $K_{k,\alpha} \leftarrow$ compute sparse α-decaying kernel with radius-based nearest neighbor search
2: $P \leftarrow$ normalize $K_{k,\alpha}$ to form a Markov transition matrix (diffusion operator; see Eq. 2)
3: $C_1, \ldots, C_M \leftarrow$ compute landmarks clusters by applying spectral clustering to P
4: $P_{NM} \leftarrow$ compute transition probabilities from points to landmarks (see Eq. 13)
5: $P_{MN} \leftarrow$ compute transition probabilities from landmarks to points (see Eq. 14)
6: $P_{MM} \leftarrow P_{MN} P_{NM}$
7: $Y_{\text{landmarks}} \leftarrow$ compute m dimensional embedding as in Alg. 1 using P_{MM} instead of P
8: $Y_{\text{points}} \leftarrow$ compute final embedding as $P_{NM} Y_{\text{landmarks}}$

The native form of PHATE as presented above is limited in scalability due to the computationally intensive step of computing potential distances between all pairs of points, as well as metric MDS. Thus, we describe here, and in Algorithm 2, an alternative way to compute a PHATE embedding that is highly scalable and provides a good approximation of the native PHATE described previously. The scalable version of PHATE uses a slight difference in computing t-step diffusion probabilities between points, it requires that every other step that the diffusion takes goes through one of a small number of "landmarks." Each landmark is selected to be a central point that is representative of a portion of the manifold, selected by spectrally clustering manifold dimensions.

First, we construct the α-decaying kernel on the entire dataset. This can be calculated efficiently and stored as a sparse matrix by using radius-based nearest neighbor searches and thresholding (i.e., setting to zero) connections between points below a specified value (e.g., 0.0001), as we regard them numerically insignificant for the constructed diffusion process. The resulting affinity matrix $K_{k,\alpha}$ will be sparse as long as α is sufficiently large (e.g., $\alpha \geq 10$) to enforce sharp decay of the captured local affinities. The full diffusion operator P is constructed from $K_{k,\alpha}$ by normalizing by row-sums as described previously.

However, powering the sparse diffusion operator would result in a dense matrix. To avoid this, we instead perform diffusion between points via a series of M landmarks where $M < N$. We select the landmarks by first applying PCA to the diffusion operator and then using k-means clustering on the principal components to partition the data into M clusters. This is a variation on spectral clustering. We then calculate the probability of transitioning in a single step from the i-th point in $\mathcal{X}$ to any point in the j-th cluster for all pairs of points and clusters. Mathematically, we can write this as

$$P_{NM}(i, j) = \sum_{\xi \in C_j} P(i, \xi) \tag{13}$$

where C_j is the set of points in the jth cluster. Thus, we can view each cluster as being represented by a landmark and the (i, j)-th entry in P_{NM} gives the probability of transitioning from the ith point in $\mathcal{X}$ to the j-th landmark. Similarly, we construct the matrix P_{MN} where the (j, i)-th entry contains the probability of transitioning from the j-th landmark to the i-th point in $\mathcal{X}$. In this case, we cannot simply sum the transition probabilities $P(\xi, i)$, $\xi \in C_j$, since we also have to consider the prior probability $Q(j, \xi)$ of the ξ-th point (with $\xi \in C_j$) being the source of a transition from a cluster C_j. For this purpose we use the prior proposed in [87], and write

$$P_{MN}(j, i) = \sum_{\xi \in C_j} Q(j, \xi) P(\xi, i) \tag{14}$$

with $Q(j, \xi) = \sum_i K_{k,\alpha}(\xi, i) / \sum_{\zeta \in C_j} \sum_i K_{k,\alpha}(\zeta, i)$.

We use the two constructed transition matrices to compute $P_{MM} = P_{MN} P_{NM}$, which provides the probability of transitioning from landmark to landmark in a random walk by walking through the full point space. Diffusion is then performed by powering the matrix P_{MM}. This can be written as

$$P_{MM}^t = P_{MN} P_{NM} P_{MN} P_{NM} \dots P_{MN} P_{NM}. \tag{15}$$

From this expression, we see that powering the matrix P_{MM} is equivalent to taking a random walk between landmarks by walking from landmarks to points and then back to landmarks t times.

We then embed the landmarks into the PHATE space by calculating the potential distances between landmarks and applying metric MDS to the potential distances. Denote the resulting embedding as $Y_{\text{landmarks}}$. We then perform an out of sample extension to all points from the landmarks by multiplying the point to landmark transition matrix P_{NM} by $Y_{\text{landmarks}}$ to get

$$Y_{\text{points}} = P_{NM} Y_{\text{landmarks}}. \tag{16}$$

Since M is chosen to vastly less than N, the memory requirements and computational demands of the powering the diffusion operator and embedding the potential distances are much lower.

The described steps are summarized in Algorithm 2. In Figure S5A-E we show that this constrained diffusion preserves distances between datapoints in the final PHATE embedding, with the scalable version giving near-identical results to the exact computation of PHATE. Further, in Figure S5B we show that the embedding achieved by this approach is robust to the number of landmarks chosen.

6.1.9 Branch Point Detection

Here we describe the methods we developed for identifying branch points and selecting representative branch- and endpoints.

Branch Point Identification We used the local intrinsic dimension estimation method derived in [21, 88] to provide suggested branch points. The procedure is as follows. Let $\mathbf{Z}_n =$

$\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ 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

$$\mathbf{L}_{\gamma,k}(\mathbf{Z}_n) = \sum_{i=1}^{n} \sum_{\mathbf{z} \in \mathcal{N}_{k,i}} ||\mathbf{z} - \mathbf{z}_i||^{\gamma}, \tag{17}$$

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,

$$\mathbf{L}_{\gamma,k}(\mathbf{Z}_n) = n^{\beta(m)} c + \epsilon_n, \tag{18}$$

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 $\beta(m)$ [88]. 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 [21, 88].

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 Figure 3A, 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:

$$\hat{\mathbf{m}}(i) = \arg\max_{\ell} \sum_{\mathbf{z}_j \in \mathcal{N}_{k,i}} \mathbb{1}(\tilde{m}(j) = \ell), \tag{19}$$

where $\mathbb{1}(\cdot)$ is the indicator function [21].

Representative Point Selection with Shake and Bake To reduce branch points and endpoints to a set of representative points for branch analysis, we use a shake and bake procedure similar to that in [23]. Let $\mathcal{V}_n = \{v_1, \ldots, v_n\}$ be the set of branch points and endpoints in the high-dimensional PHATE coordinates that we wish to reduce. We create a Voronoi partitioning of these points as follows. We first permute the order of $\mathcal{V}_n$, which we denote as $\mathcal{V}'_n = \{v_{1'}, \ldots, v_{n'}\}$. We then take the first point $v_{1'}$ and find all the points in $\mathcal{V}'_n$ that are within a distance of h, where h is a scale parameter provided by the user. These points (including $v_{1'}$) are assigned to the first component of the partition and removed from the set $\mathcal{V}'_n$. This process is then repeated until all points in V_n are assigned to the partition. To ensure that each point is assigned to the nearest component of the partition (as measured by proximity to the centroid), we next calculate the distance of each point to all centroids of the partition, and reassign the point to the component with the nearest centroid. This reassignment process is repeated until a stable partition is achieved. This completes the process of constructing the Voronoi partition.

The Voronoi partition constructed from this process may be sensitive to the ordering of the points in $\mathcal{V}_n'$. To reduce this sensitivity, we repeat this process multiple times (e.g., 40-100) to create multiple Voronoi partitions. We then construct a distance between points by estimating the probability that two points are not in the same component from this partitioning process. This provides a notion of distance that is robust to noise, random permutations, and the scale parameter h. We then partition the data again using the above procedure except we use these probability-based distances. The representative points are then selected from the resulting centroids of this final partition.

Branch Detection We now describe how we assign data points to branches using the correlation and anticorrelation of neighborhood distances (in higher dimensional PHATE coordinates). The approach is demonstrated visually in Figure 3Aiii. Here we consider two reference cells X and Y. We wish to determine if cells $Q1$ and $Q2$ belong to the branch between X and Y or not. Consider $Q1$ first which does belong to this branch. If we move from $Q1$ towards X, we also move farther away from Y. Thus the distances to X and Y of a neighborhood of points around $Q1$ (which will be located on the branch) are negatively correlated with each other. Now consider $Q2$ which does not belong to the branch between X and Y. In this case, if we move from $Q2$ towards Y, we also move closer to X. Thus the distances to X and Y of a neighborhood of points around $Q2$ are positively correlated with each other. In practice, these distance-based correlations are computed for each possible branch and the point is assigned to the branch with the largest anticorrelation (i.e. the most negative correlation coefficient).

6.1.10 Divergence-Based Analysis

Here we discuss the divergence-based methods we used to analyze the EB scRNA-seq data. We first discuss the use of the KL divergence for testing statistical significance between regions of correlation coefficients. We then describe the EMD-based score analysis which was used to perform differential expression analysis within the EB scRNA-seq data to identify genes uniquely associated with different parts of the PHATE visualization.

Statistical Significance Based on KL Divergence The KL divergence was used to test for statistical significance between groups of correlation coefficients when comparing the scRNA-seq EB data with bulk RNA-seq samples. The KL divergence is a measure of difference between two probability distributions and is zero if and only if the two distributions are identical almost everywhere [83]. It is defined between two probability densities p and q as

$$D_{KL}(p||q) = \int p(x) \log \left(\frac{p(x)}{q(x)} \right) dx. \tag{20}$$

In our case, we compare the distributions of correlation coefficients from each pair of regions defined in the EB scRNA-seq data. The KL divergence is estimated using the nonparametric k-nn estimator in [89]. A central limit theorem exists for this estimator [90], from which we

construct a p-value on the hypothesis that the KL divergence is zero (i.e. the distributions of correlation coefficients from the two regions are identical). This is a generalization of the mean comparison test where we compare the overall shapes of the two distributions instead of just the means and is a primary reason we chose the KL divergence for this task over other measures (such as the EMD).

EMD Score Analysis The EMD is another measure of dissimilarity between two probability distributions that is particularly popular in computer vision [61]. The EMD was chosen to perform differential expression analysis in the EB scRNA-seq data due to its stability in estimation compared to other divergence measures. Intuitively, if each distribution is viewed as a pile of dirt, the EMD can be thought of as the minimum cost of converting one pile of dirt into the other. If the distributions are identical, then the cost is zero. When comparing univariate distributions (as we do as we only consider a single gene at a time), the EMD simplifies to the L^1 distance between the cumulative distribution functions [91]. That is, if P and Q are the cumulative distributions of densities p and q, respectively, then the EMD between p and q is $\int |P(x) - Q(x)|dx$. While the EMD is nonnegative, we assign a sign to the EMD score based on the difference between the medians of the distributions.

6.2 Comparison of PHATE to Other Methods

We compare PHATE to methods of dimensionality reduction, graph rendering, and visualization on several datasets in Figures 4, 5, and S1. The datasets used in the comparisons include (in the order shown in Figure S1): A. Artificial tree data with 7 branches, at low, medium and high levels of noise; B. DLA fractal tree; C. Three intersecting curves; D. Video of a rotating teapot [92]; E. Swiss roll data used in [6]; F. Frey faces video [28]; G. Developing mouse bone marrow cells, enriched for the myeloid and erythroid lineages, which were measured with the MARS-seq single cell RNA-sequencing technology [16]; H. Single cell RNA-sequencing of epithelial cells from mouse small intestine and organoids [93]; I. Mass cytometry data measuring T cell development into CD8+ and CD4+ T cells in mouse thymus [2]; J. New embryoid body data from a 27-day timecourse; K. Mass cytometry data showing iPSC reprogramming of mouse embryonic fibroblasts [14]; L. Single cell RNA-sequencing of mouse retinal bipolar cells [13]; M. Single cell RNA-sequencing of mouse cortical cells from the somatosensory cortex and hippocampal CA1 region [25]; N. Gaussian mixture model with two touching clusters and two disjoint clusters; O. Columbia Object Image Library (COIL-20), 20 videos of rotating objects [94]; P. MNIST, 70,000 images of handwritten digits from 0 to 9 [95]. Some of these biological datasets represent differentiating processes within the body, and hence visualizing progression is key to understanding the structure of these datasets. Other artificial datasets show a combination of clusters and trajectories, while the artificial datasets give a plausible range of manifold structures that could be found in biological data.

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, t-SNE, and diffusion maps. However, because PHATE can be used to extract trajectory or differentiation structure, we also consider tools that find and *render* explicit "differentiation tree structures"; these methods include SPADE [96, 97] and Monocle2 [4].

Finally, we note that several methods exist that focus on finding *pseudotime* orderings of cells, such as Wanderlust [17], Wishbone [2], and diffusion pseudotime [3]. Wanderlust can find single non-branching progressions. Wishbone recognizes a single branch, while diffusion pseudotimes provides potentially multiple branches. These pseudotime methods can be used alongside PHATE to order parts of the branching progressions. Indeed, we use Wanderlust to extract ordering from the branches identified from PHATE.

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.

Note that we do not include any meta data, such as sample time or clusters, in any of the analyses. Therefore, we are focusing on the performance of these methods in an unsupervised setting. Instead, we use this meta data as a tool for comparing the results of the various methods.

Comparison of PHATE to Dimensionality Reduction Methods

Figures 4 and 5 compare the PHATE visualization to the dimensionality reduction methods of principal components analysis (PCA), Diffusion Maps (DM), TSNE, Isomap, and UMAP on four different artificial datasets and five different biological datasets. Further comparisons to Multidimensional Scaling (MDS), MDS on Diffusion Maps, t-SNE on Diffusion Maps, Locally Linear Embedding (LLE) and all above mentioned methods on a total of eighteen real and artificial datasets are shown in Figure S1. The datasets show a combination of intersecting and distinct manifolds, clusters, and branching trajectories, examining a range of both real biological manifolds and plausible challenging structures for visualization. For all datasets, the PHATE visualization is best at distinguishing branches and trajectories and discovering the underlying structure of the data. We focus on each method individually.

Comparison of PHATE to PCA: PCA is a popular method of data analysis that uses eigen-decomposition of the covariance matrix to learn axes within the high-dimensional data that account for the largest amount of variance within the data [98]. However, PCA assumes a linear structure on the data, the visualization amounts to projecting the data onto a slicing plane, which creates a noisy visualization. Also, since biological data are rarely linear, PCA is unable to optimally reduce non-linear noise along the manifold and reveal progression structure in low dimensions. This is evident in Figure 4A and B where we compare PCA to PHATE on artificial tree data. This data contains seven distinct branches uniformly sampled in 60 dimensions. See

Section 6.4.2 for details. PCA does capture some of the global structure in this relatively low-noise data. However, many branches are not visible in the first two PCA dimensions and the trajectories in the PCA visualization are noisy compared to the PHATE visualization, in which all seven branches are easily identifiable.

For the other datasets in Figures 4 and 5, PCA captures some of the overall global structure of the datasets. For example, the PCA dimensions in Figure 5D encode the overall time progression of the noisy EB scRNA-seq data. Thus, PCA captures some of the global structure. However, PCA presents mostly a cloud of cells in this case and any finer branching structure is not visible. This contrasts with PHATE which shows multiple branches and trajectories. Similar results are obtained from the mouse bone marrow scRNA-seq and iPSC CyTOF datasets in Figures 5A and E respectively, demonstrating that PCA is unable to accurately visualize the global and local structure of the data simultaneously.

Comparison of PHATE to t-SNE: t-SNE (t-distributed stochastic neighbor embedding) [5] is a visualization method that emphasizes local neighborhood structure within data. Recently, t-SNE has become popular for revealing cluster structure or separations in single cell data [99]. However, due to its emphasis on preserving local neighborhoods, t-SNE tends to shatter trajectories into clusters as seen in the artificial tree and intersecting curves in Figures 4A and C as well as the EB data and the iPSC data in Figures 5D, E, and S1D. In all of these cases, the data naturally have a strong trajectory structure either by design (the artificial trees and intersecting curves) or due to the developmental nature of the data (the EB and iPSC datasets). Thus t-SNE creates the false impression that the data contain natural clusters, which could lead to incorrect analysis.

Furthermore, the adaptive kernel used in t-SNE for calculating neighborhood probabilities tends to spread out neighbors such that dense clusters occupy proportionally more space in the visualization compared to sparse clusters [100]. Thus, the relative location of data points within the t-SNE embedding often does not accurately reflect the relationships between them. This is clearly visible in the t-SNE plot in Figure 4A where the shattered branches are located far away from where they originated in the main structure in the artificial tree data. Similarly, t-SNE creates clusters in the EB and iPSC data in Figures 5D and E which split the time samples into different components. Since the relative position of clusters in t-SNE is generally meaningless the overall progression of the data is destroyed.

Even in the case where the data are more naturally separated into clusters, t-SNE can destroy the global information about the relative relationships between clusters due to this weakness. In contrast, PHATE separates clusters that are sufficiently separated from each other (see Figure 4D) while maintaining the relative relationships of clusters based on the relative positions of the clusters in the PHATE embedding. In other words, PHATE preserves both the global and local structure while t-SNE only preserves local structure.

One proposed solution to the failure of t-SNE to retain global structure is to use a random walk to learn the global structure, and then apply t-SNE to the resulting kernel [5]. One approach to do this is to apply t-SNE to DM. However, our experiments show that this fails to

capture the global structure of the data. In the artificial tree data (Figure S1A and B) the intersecting curves (Figure S1E), as well as the new EB data (Figure S1L) t-SNE on DM shatters trajectories. Furthermore, in the retinal bipolar dataset (Figure S1N) t-SNE on DM shatters clusters and creates misleading trajectory-like structures. Hence, performing t-SNE on diffusion maps suffers from the same shortcomings as t-SNE, with additional distortion from the denoising aspect of diffusion maps, as t-SNE tends to perform better on noisier data (see Figures 4A and B). Due to the nature of the t-SNE penalty function, global distances encoded in the diffusion distances are ignored, and the resulting embedding is a denoised equivalent of t-SNE, which is more prone to shattering trajectories than t-SNE and lacks the global structure from DM.

Comparison of PHATE to Diffusion Maps: Diffusion maps effectively encode continuous relationships between cells. However, different trajectories are often encoded in different dimensions (i.e., since they represent different meta-stable states of the diffusion process) as seen in Figure S2B, which is unsuitable for visualization. In contrast, PHATE effectively encodes trajectories in lower dimensions for visualization. This is also seen clearly in Figures 4A and C in the comparison of PHATE to diffusion maps on the artificial tree and intersecting curves data (for each data set, the same kernel and diffusion scale t is used for both diffusion maps and PHATE). In this case, the diffusion maps visualization is denoised and the global structure is visible. However, multiple branches are not visible in the low-dimensional visualization of diffusion maps. In fact, approximately six diffusion maps coordinates are needed to separate all ten branches (see Figure S2B). In contrast, all of the ten branches of the artificial tree data are clearly visible in the PHATE visualization. Similarly, multiple branches that are visible in the PHATE visualization are not visible in the diffusion maps visualization for the bone marrow and EB scRNA-seq datasets (Figures 5A and D, respectively). Additionally, the diffusion maps instabilities mentioned previously appear to cause very noisy data (e.g. the scRNA-seq data in Figures 5A,B, and D) to contract too much into thin trajectories, which can distort some of the underlying progression structure. In summary, while diffusion maps works well for nonlinear dimensionality reduction, it is not well-suited for visualizing data with multiple trajectories due to its instabilities and its propensity to encode different trajectories in different dimensions.

A logical question is whether applying MDS on diffusion distances would be sufficient for encoding the high dimensional spatial information from diffusion maps in low dimensions for visualization. However, in Figures S1E and Q on the intersecting curves and COIL20 we show that MDS on DM suffers equivalently from instability at boundary conditions and intersections of manifolds, producing a totally structureless embedding. Additionally, MDS on DM collapses trajectories into thin trajectories as shown on the mouse bone marrow scRNAseq, new EB data, and iPSC CyTOF data (Figures S1I, L, and M respectively), distorting the intricate structure visible in PHATE. Thus, performing MDS on diffusion maps, without applying the informational potential transformation of PHATE, is insufficient for high quality visualization.

Comparison of PHATE to MDS Multidimensional scaling (MDS) [101] aims to preserve or approximate the metric structure of the data by optimizing a stress loss between original (Euclidean) distances and embedded ones. While MDS ostensibly preserves both local and global distances, it does not rely on or infer any particular intrinsic structure from the data. Therefore, it cannot separate meaningful relations in complex high dimensional data from superfluous ones. In particular, this causes MDS to be strongly affected by noise, as shown in the artificial trees in Figures S1C and D where the embeddings show no structure at all. Further, this causes problems in visualizing noisy biological datasets, such as the new EB scRNAseq data (Figure S1L) where local trajectories are lost, and the iPSC CyTOF data (Figure S1M) where the branching structure is entirely masked by noise. MDS also fails to separate clusters in noisy data, as shown in biological datasets from [25] and [13] (Figures S1N and O) as well as in MNIST (Figure S1R).

Comparison of PHATE to Isomap: Isomap [6] embeds the intrnisic metric structure of the data by applying MDS to geodesic distances, which are obtained by constructing a k-nearest neighbor graph over the data, and then applying all-pairs shortest path search (e.g., Dijkstra [102] or Floyd [103]) to compute distances. Like other manifold learning methods, Isomap works under the assumption that the data is sampled from an underlying manifold, and thus the geodesic distances approximate intrinisic manifold distances and the coordinate assigned by MDS should provide a global intrinsic coordinate system. However, this assumption is mainly valid when the manifold itself is convex, with no holes, and the data is sampled uniformly from it with only small amount of noise away from the manifold. Indeed, the main weaknesses of Isomap is its topological instability when such assumptions are not satisfied in practice [104–106], as we also show here. These instabilities render Isomap susceptible to spurious connections created in noisy datasets, as shown by the failure to separate branches on the DLA Tree (Figure S1D). Further, Isomap is incapable of embedding clusters, such as the Gaussian Mixture Model (Figure 4D), the mouse retinal bipolar scRNAseq (Figure 5B), and MNIST (Figure S1R), in which many clusters are merged together since geodesic distances do not consider the data distribution and do not quantify relations between disconnected clusters. Additionally, Isomap is also unstable to intersecting manifolds, as shown by the Intersecting Curves dataset, where Isomap fails to distinguish between two of the curves and shows no intersections (Figure 4C), and does not clearly display branching points, as shown on the iPSC data where Isomap only separates points by timepoint (Figure 5E).

Comparison of PHATE to Locally Linear Embedding: Local Linear Embedding (LLE) [28] is a manifold learning algorithm that is similar to Isomap in that it assumes the data is sampled from a single smooth manifold, approximated by a k-NN graph, and tries to use its geometric properties to embed the data. However, unlike Isomap, it only considers local information - namely, it uses the low dimensional coordinate neighborhoods, which are (independently) linearly related to local manifold patches, and tries to tile these local coordinates into a consistent global embedding in low dimensions. This is done by first optimizing weights that allow the

approximation of each data point as a linear combination of its neighbors, and then optimizing low dimensional coordinates that preserve the same linear relations encoded by these weights.

While LLE is less susceptible to shortcut connections, it heavily relies on a smooth well-connected manifold structure to provide global connections through the data. Therefore, LLE is ill-suited for embedding separate clusters, as shown by its failure to separate the distinct digits in the MNIST dataset (Figure S1R). Further, dure to its reliance on local interpolation for encoding relations between each point and its neighbors, it strongly affected by boundary conditions that make such interpolation unstable, as shown by its complete failure to embed the Gaussian Mixture Model (Figure S1P). Additionally, LLE does not handle intersecting manifolds and high curvatures, due its implicit assumption that local data patches correspond to manifold coordinate neighborhoods, which can be linearly approximated by a tangent space of the same intrinsic dimension of the manifold. For example, on the Intersecting Curves dataset, while PHATE shows a clean intersection between the curves, LLE treats one of the intersections as a noisy subcluster, and the other intersection is not shown, with the orange curve shown simply as a continuation of the blue curve (Figure S1E). Finally, we not that some additional weaknesses of LLE for visualization of even simple syntheric biomedical data were reported previously, e.g., in [104, 107].

Comparison of PHATE to UMAP Like t-SNE, upon which the UMAP algorithm is modeled, UMAP encourages formation of clusters even when cluster structures do not exist. As such, UMAP shatters trajectories, as shown in the artificial trees and intersecting circles (Figures 4A and C respectively,) the new EB data (Figure 5D) and most obviously on the DLA Tree, where UMAP splits the tree into two separate trees and one lone disjoint cluster. (Figure S1D). Additionally, although UMAP's algorithm is designed to respect global relationships between the clusters that it forms, this does not always hold between distant clusters; take for example the Gaussian Mixture, in which UMAP swaps the relationship between the green and blue clusters (green should be closer to blue than to orange) and places the blue and orange clusters artificially far apart (Figure 4D).

Comparison of PHATE to Graph-Rendering Methods

Graph-rendering methods differ fundamentally from dimensionality reduction methods in that they do not produce a reduced dimension representation of the data and instead focus only on providing a specific rendering of the data. However, these renderings are often limited by structural assumptions on the data (e.g. a tree) that may be inaccurate. Figures 4 and 5 compare the PHATE visualization to graph-rending methods including Force Directed Layout (FDL) and Monocle2.

Comparison of PHATE to Force-Directed Methods: FDL algorithms attempt to draw a weighted graph in a two-dimensional space so that all edge weights are approximately preserved as distances and relatively few edges cross each other. To this end, such methods solve a n-body

problem modeled by attractive and repulsive forces that resemble physical systems, such as elastic (e.g., Hooke's law), electric (e.g., Coulomb's law), or nuclear forces. Typically, repulsive forces are used to separate and spread out all pairs of nodes while attractive ones are used to keep neighboring nodes on the graph close to each other in the embedding. These attractive forces are scaled based on the strength of the connections within the graph. We specifically apply the Spring Layout method within the NetworkX Python package [108] to the graph defined by the α-decaying kernel produced for PHATE. This method uses the Fruchterman-Reingold algorithm [109], which is motivated by the aesthetics of rendering a planar graph, and models the attractive and repulsive forces based on a combination of notions elastic attraction, nuclear repulsion, and global stability criteria (e.g., ideal uniform distance and decaying temperature of the system).

FDL algorithms are generally computationally expensive, making it difficult to scale them to larger datasets (Figure S5E). FDL algorithms also do not denoise the connections between data points, but rather assume that attractive and repulsive forces will eventually balance each other. Thus, they suffer from the same sensitivity to graph construction as Isomap, where in this case spurious connections in noisy data can dominate the resulting force-dynamics and strongly affect the embedding (see the artificial tree and intersecting curves in Figures 4A and C respectively). Additionally, exceedingly weak forces between distant clusters, which are clearly not well-connected in the graph, lead to a failure to retain long-range global distances, as shown in Figure 4D. These shortcomings can lead to the loss of information, such as a lack of branching structure in the iPSC CyTOF data, or a merging or clusters in the neuronal scRNAseq from [25] (Figures 5E and C respectively.)

Comparison of PHATE to Tree-Rendering Methods (SPADE and Monocle2): SPADE [96, 97] and Monocle2 [4] are popular methods that fit the data to a predetermined structure such as a tree. These methods first attempt to do data reduction by clustering the data. 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. Additionally, the results from these methods will be incorrect if the underlying data does not lie on a tree. In contrast, PHATE does not make any assumptions on the data and instead learns the underlying structure.

SPADE fits a minimal spanning tree to the clusters and was originally designed for mass cytometry data [96]. In Anchang et al. [97], the authors applied SPADE to scRNA-seq data by selecting relevant genes to perform dimensionality reduction. This makes it difficult to do data exploration as gene selection must be performed first. In contrast, PHATE does not require any gene selection procedure although PHATE can be used to analyze specific genes of interest by including only the relevant genes. SPADE has several other limitations according to Anchang et al. [97]. First, the SPADE results can be sensitive to the number of clusters which must be specified by the user. Second, down-sampling is required to visualize large datasets. SPADE is

very sensitive to random down-sampling and will produce very different trees even when only down-sampling to 99% [97]. Thus as with scTDA, the random nature of the SPADE results make it difficult to discover the right structure of the data with SPADE. Given these limitations, we do not make a direct comparison between PHATE and SPADE.

Monocle2 also fits a tree to cell clusters using the DDRTree algorithm [110, 111] as a default. We compare Monocle2 to PHATE in Figures 4, 5 and S1. For the artificial tree data in Figures 4A and B, Monocle2 fails to detect multiple branches in the high noise setting. For the bone marrow scRNA-seq data in Figure 5A, Monocle2 fails to detect several branches that are visible in the PHATE visualization. At the same time, Monocle2 shows several branches that are not detected by PHATE. However, the number and location of these branches vary from run to run on the same data with the same settings. Thus it is difficult to determine if the branches shown by Monocle2 are spurious or not.

Similar results are obtained when applying Monocle2 to the EB scRNA-seq data where the number and location of the branches within the Monocle2 visualization differ drastically from run to run. Thus it is difficult to determine the underlying structure of this data using Monocle2. In contrast, for the same set of parameters, PHATE produces the same results with each run while preserving the relative relationships between different branches directly in the visualization based on their proximity.

6.3 PHATE for Data Exploration on Non-Single-Cell Data

6.3.1 PHATE on High-dimensional High-throughput Data

As a general dimensionality reduction method, PHATE is applicable to many datatypes. Here we show that PHATE reveals and preserves global transitional structure in microbiome data, human SNP data, and (non-biological) image data.

PHATE on Microbiome Data Reveals Archetypal Structure Recently there have been many studies of bacterial species abundance in the human intestinal tract, saliva, vagina and other membranes as measured by sequencing of the 16S ribosomal-RNA-encoding gene (16s sequencing) or by whole genome shotgun sequencing (metagenomics). However, most analysis of microbiome data has been limited to clustering and PCA. Here we use PHATE to analyze microbiome data.

First we note that PCA (Figure 6A 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 S6A shows the PHATE embedding colored by two genera (bacteroides and prevotella) and a phylum (actinobacteria) of bacteria on the same 9660 samples as in Figure 6A. 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 [112] 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 Figure S6B, we see 4 major branches, instead of the three enterotypes reported in previous literature [113], 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 typically seen in archetypal analysis [114, 115]. 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.

PHATE on SNP Data Reveals Geographic Structure To demonstrate PHATE on population data, we examined a dataset containing 2345 present-day humans from 203 populations genotyped at 594,924 autosomal SNPs with the Human Origins array [27]. In Figure 6B, we see that as compared to PCA, 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 show as orange diamonds (see Figure 6G for population labels.) 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. PCA tends to crowd populations together into two linear branches, without clearly distinguishing between population groups or showing population divergence.

PHATE on Facial Images To demonstrate that PHATE can also be used to learn and visualize the underlying structure of nonbiological data, we applied PHATE to the Frey Face dataset used in [28]. This dataset consists of nearly 2000 video frames of a single subject's face in various poses. Figure 6C shows a 3D visualization of this dataset using PHATE, colored by time. Multiple branches are clearly visible in the visualization and each branch corresponds to the progression of a different pose. A short video highlighting two of these branches is available at `https://www.youtube.com/watch?v=QMCWsKNgvHI`. The video highlights the continuous nature of the data as points along branches correspond to transitions from pose to pose.

6.3.2 PHATE on Connectivity Data

Thus far we have used PHATE to embed high-dimensional data. That is, we have mapped the original feature space of the data to the PHATE dimensions. However, the PHATE algorithm also allows for embedding any data that exists in either an inner product space or a metric space. In other words, we can apply PHATE to data that is naturally described by distances or affinities

instead of features. For example, network data (i.e. graphs) can be embedded using PHATE simply by skipping the initial steps of the algorithm that go from the original feature space to an affinity matrix. We can thus replace the affinity matrix with the network (as represented by an affinity matrix) and proceed with the rest of the PHATE algorithm.

There are abundant examples of natively networked data in biology, such as chromatin conformation contact maps (Hi-C), gene/protein interaction networks, and neural connectivity data such as fMRI. Outside of biology, network data is also prevalent. A common example is social network data that contain information of friend-communities (e.g. Facebook) or interest-communities (e.g. Twitter). We show that PHATE provides a visualization of network data that emphasizes major structure (i.e. pathways) in the network, better than typical graph layout methods.

PHATE on Hi-C Data Reveals Spatial Chromatin Structure We use PHATE to visualize human Hi-C data from [29] by using the Hi-C contact map as the affinity matrix in the PHATE algorithm. The contact map gives the frequency with which genomic locations are observed in spatial proximity. Hi-C contact maps are typically visualized using the matrix directly – often using the 45 degree counterclockwise rotated upper triangle part of the matrix [29]. While this depiction can show chromosomal domains, it is not a reconstruction of the actual spatial structure of the chromatin. In contrast, the PHATE embedding reconstructs the relative positions of the genomic locations both locally and globally in such a way that the embedding represents an actual projection of the spatial structure within and between chromosomes. As a result, we get an intuitive visual of the Hi-C contact map that not only shows the topological domains present in the data but also how they are connected to one another. Figure 6D shows a 3D PHATE embedding of the chromatin, colored by chromosome (see Figure S6C for a 2D embedding of the same data). We see that the embedding resembles the fractal globule structure proposed in [116], with the total chromatin organized in a spherical shape and individual chromosomes mostly connected within themselves.

PHATE can also effectively visualize a single chromosome. Figure 6E shows PHATE just on chromosome 1 contact map at 10 kilobase (kb) resolution. A 3D PHATE visualization is given in Figure 6E, left. Each point corresponds to a genomic fragment and is colored by its location within the genome. In this visualization, multiple "folds" are clearly visible.

To validate that the PHATE embedding of Hi-C data is meaningful we color the embedding by the ChIP-seq signals of several chromatin modification markers. Figure S6C shows a 2D PHATE embedding of chromosome 1 colored by various methylation and acetylation markers (ChIP-seq [117], dataset ENCSR977QPF). Histone methylation and acetylation play an important role in global gene regulation via control of chromatin organization. They are often used in combination with Hi-C data to investigate the open or closed structure of chromatin [29], where the ChIP-seq signal of various histone modification markers correlates with so called topologically associated domains or TADs. Figures S6C and 6E show an organized methylation (H3K27me3 and H3K4me2) and acetylation (H3K9ac and H3K27ac) signal on PHATE, with clusters of similar intensity of the marker, suggesting that the PHATE embedding has biological

meaning with respect to the spatial organization of the chromatin.

These results indicate that PHATE provides a visualization of Hi-C data that captures more spatial information compared to directly looking at the contact map.

PHATE on Facebook Data Reveals Super Connectors We visualize Facebook network data using PHATE. The data [118] consist of networks of friends. This network graph is directly converted into a 0-1 affinity matrix and fed to PHATE. Figure 6F compares the PHATE visualization of the network to a force-directed layout, a common method for visualizing network data. In both plots, edges (friends) are shown and each node/person is colored by degree (the number of friends a person has). The PHATE embedding clearly shows multiple branching structures in the network, that are not visible in the force-directed layout which shows a T shape.

Several subnetworks and important nodes (super-connectors) between subnetworks are visible in the PHATE embedding that are not visible in the force-directed layout visualization. For example, in the top-left corner of the PHATE visualization, a single node connects the top-left group of people to the remainder of the network. Several other important nodes can be identified that bridge the gap between the left section of the network and the center region. Thus PHATE can be used for visualizing network data and identifying important features of the network.

We also show that PHATE can be used to find more structure within subnetworks as identified by the friend networks of selected individuals (referred to as ego nodes in [118]) in Figure S6D. Again, we find that PHATE finds more structure in subnetworks.

Therefore, we see that PHATE can be used to visualize any type of data, high-dimensional or featureless. Further, we see that PHATE will emphasize continuous transitional structure, while maintaining separation between clusters in these datasets.

6.4 Experimental Methods

The processes for generating the EB data and the artificial tree data are described in this section.

6.4.1 Generation of Human Embryoid Body 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 (Figure S7H). 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.

6.4.2 Construction of the Artificial Tree Test Case

The artificial tree data shown in Figure 1B is constructed as follows. The first branch consists of 100 linearly spaced points that progress in the first four dimensions. All other dimensions are set to zero. The 100 points in the second branch are constant in the first four dimensions with a constant value equal to the endpoint of the first branch. The next four dimensions then progress linearly in this branch while all other dimensions are set to zero. The third branch is constructed similarly except the progression occurs in dimensions 9-12 instead of dimensions 5-8. All remaining branches are constructed similarly with some variation in the length of the branches. We then add 40 points at each endpoint and branch point and add zero mean Gaussian noise with a standard deviation of 7. This construction models a system where progression along a branch corresponds to an increase in gene expression in several genes. Prior to adding noise, we also constructed a small gap between the first branch point and the orange branch that splits into a blue and purple branch (see the top set of branches in the left part of Figure 1B). This simulates gaps that are often present in measured biological data.

6.5 Data Preprocessing

In this section, we discuss methods we used to preprocess the various datasets.

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 [1].

Single-Cell RNA-Sequencing Data Preprocessing This data was processed from raw reads to molecule counts using the Cell Ranger pipeline [119] Additionally, to minimize the effects of experimental artifacts on our analysis, we preprocess the scRNAseq data. We first filter out dead cells by removing cells that have high expression levels in mitochondrial DNA. In the case of the EB data which had a wide variation in library size, we then remove cells that are either below the 20th percentile or above the 80th percentile in library size. scRNA-seq 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

[120]. Removing cells with extreme library size values helps to correct for these technical variations. We then drop genes that are only expressed in a few cells and then perform library size normalization. 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 take the square root transform of the data and then 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.

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 scRNA-seq data. We first perform "library size" normalization to account for technical variations in different samples. We then log transform the data and then use PCA to reduce the data to 30 dimensions.

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 that are shown in Figure 1D.

ChIP-seq Processing for Hi-C Visualization We used narrow peak bed files and took the average peak intensity for each bin at a 10 kb resolution. For visualization, we smoothed the average peak intensity values based on location using a 25 bin moving average.

6.6 Software

Python, R and Matlab implementations of PHATE are available on GitHub, for academic use:

`https://github.com/KrishnaswamyLab/PHATE`

An interactive tool for exploring PHATE on the EB data is available at:

`https://www.krishnaswamylab.org/phatewebtool`