Visualizing Transitions and Structure for High Dimensional Data Exploration

Kevin R. Moon,[1,2,3†] David van Dijk,[1,3†] Zheng Wang,[4†] Daniel Burkhardt,[1]
William S. Chen,[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

In the era of '*Big Data*' there is a pressing need for tools that provide human interpretable visualizations of emergent patterns in high-throughput high-dimensional data. Further, to enable insightful data exploration, such visualizations should faithfully capture and emphasize emergent structures and patterns without enforcing prior assumptions on the shape or form of the data. In this paper, we present PHATE (Potential of Heat-diffusion for Affinity-based Transition Embedding) - an unsupervised low-dimensional embedding for visualization of data that is aimed at solving these issues. Unlike previous methods that are commonly used for visualization, such as PCA and tSNE, PHATE is able to capture and highlight both local and global structure in the data. In particular, in addition to clustering patterns, PHATE also uncovers and emphasizes progression and transitions (when they exist) in the data, which are often missed in other visualization-capable methods. Such

patterns are especially important in biological data that contain, for example, single-cell phenotypes at different phases of differentiation, patients at different stages of disease progression, and gut microbial compositions that vary gradually between individuals, even of the same enterotype.

The embedding provided by PHATE is based on a novel informational distance that captures long-range nonlinear relations in the data by computing energy potentials of data-adaptive diffusion processes. We demonstrate the effectiveness of the produced visualization in revealing insights on a wide variety of biomedical data, including single-cell RNA-sequencing, mass cytometry, gut microbiome sequencing, human SNP data, Hi-C data, as well as non-biomedical data, such as facebook network and facial image data. In order to validate the capability of PHATE to enable exploratory analysis, we generate a new dataset of 31,000 single-cells from a human embryoid body differentiation system. Here, PHATE provides a comprehensive picture of the differentiation process, while visualizing major and minor branching trajectories in the data. We validate that all known cell types are recapitulated in the PHATE embedding in proper organization. Furthermore, the global picture of the system offered by PHATE allows us to connect parts of the developmental progression and characterize novel regulators associated with developmental lineages.

1 Introduction

High dimensional, high-throughput data is accumulating at a staggering rate nowadays in all fields of science and technology, including process monitoring, cybersecurity, finance, social networking, biology, and health. The massive volumes of collected data require automatic tools that are able to process and organize them into human-interpretable representations. Further, such tools are expected and required to work in unsupervised data-driven settings in order to enable data exploration that supports initial knowledge discovery and hypothesis forming.

In particular, there is an increasing demand in modern exploratory data analysis for visualization of emergent patterns and structures in big high dimensional data. However, this demand is met by a paucity of methods that are clearly designed for the purpose of visualizing such data to understand both the global and local structure without requiring a priori model fitting. To the best of our knowledge, the only currently popular candidate for such a visualization is t-distributed Stochastic Neighborhood Embedding (tSNE) [1]. We note that, by design, tSNE is mainly effective in revealing and emphasizing cluster structures in data. While these structures are important, they are not the only sought after structures in data exploration. Indeed, Big Data often contain dynamical changes exhibited by transitional structure between parts of the data space. These are not well modeled by separated clusters (as shown in Figure 1B), and are better represented in the form of progression pathways or transitions. Therefore, to capture and visualize progression structures, we present a new embedding, ***PHATE (Potential of Heat-diffusion for Affinity-based Transition Embedding)***. PHATE is geared towards optimally visualizing, in a clean, denoised, and robust fashion, cohesive global nonlinear and local structures in big high-dimensional data, including nonlinear pathways and progressions. In par-

ticular, we demonstrate via extensive validation that PHATE is especially useful in visualizing big high-dimensional biomedical data.

Biological data often contain a continuous spectrum of cells that create progressions (see Figure 1A). Progression is inherent to single-cell data since all human cells arise from a single oocyte that differentiates into the various tissues and subtypes. Progression is also inherent to other biological data types. For example, gut bacterial species in patients with autoimmune conditions can show progression based on the extent of the underlying disease. Population genetic data can show progression in genotypes based on population drift and admixture events. Snapshot data from these particular systems will show evidence of continuous progression and trajectories. In addition, feature-less data with only connectivity measurements, such as Hi-C contact maps or neural connectivity data, can be visualized to highlight major progressions. However, the high-dimensional and noisy nature of these data makes it difficult to extract or visualize the progression (see Figure 1A) or to use it for data exploration. PHATE can visualize any of the above (and more types of data) by using a new type of distance between data points that emphasizes connectivity through data and eliminates noise.

Unlike many recent methods in computational biology that attempt to discover pseudotime trajectories, PHATE does not impose any type of structure on the data and does not require user-supervision. As such it is vastly different from methods such as Monocle [2,3], Wishbone [4], or Wanderlust [5], which impose specific structures or orderings on the data. In contrast, PHATE preserves the inherent structure and connectivity of the data including major and minor pathways and branches or bifurcations as well as separate clusters. PHATE visualizations enable researchers to better understand and identify data connectivity, branch points, and local intrinsic dimensionality of the data while simultaneously cleaning the data via PHATE's internal diffusion process. Further, PHATE allows for an intermediate higher dimensional embedding for the purposes of clustering and branch analysis. Also, methods that extract pseudo-time orderings from data, such as Wanderlust and Wishbone, can be run on the higher-dimensional PHATE embedding once the structure has been identified from the visualization.

We validate PHATE on a variety of datasets to demonstrate its versatility and ability to produce insights in many types of systems and measurement modalities, including non-biological data and data that are natively described as a network. Among these are simulated data with multiple branches, single-cell RNA-sequencing (scRNA-seq) data, mass cytometry (CyTOF) data, microbiome data, single-nucleotide polymorphism (SNP) population genetics data, facial images, Facebook network data, and Hi-C chromatin interaction data.

To demonstrate the importance of the structure revealed by PHATE, we validate PHATE on a newly generated single-cell RNA-sequencing dataset measured from embryoid body cells collected in 3 day increments over a period of 27 days. From the PHATE visualization, we identify and analyze the cells in multiple branches, including embryonic stem cells (ESC), neural crest cells (NCC), neural progenitors, cardiac progenitors, and cells from the mesoderm and ectoderm layers (see Figure 1C). We validate these findings by comparing the scRNA-seq expression levels with bulk RNA-seq expression levels measured from the various cell types.

Finally, we describe methods for extracting quantitative information from PHATE such as

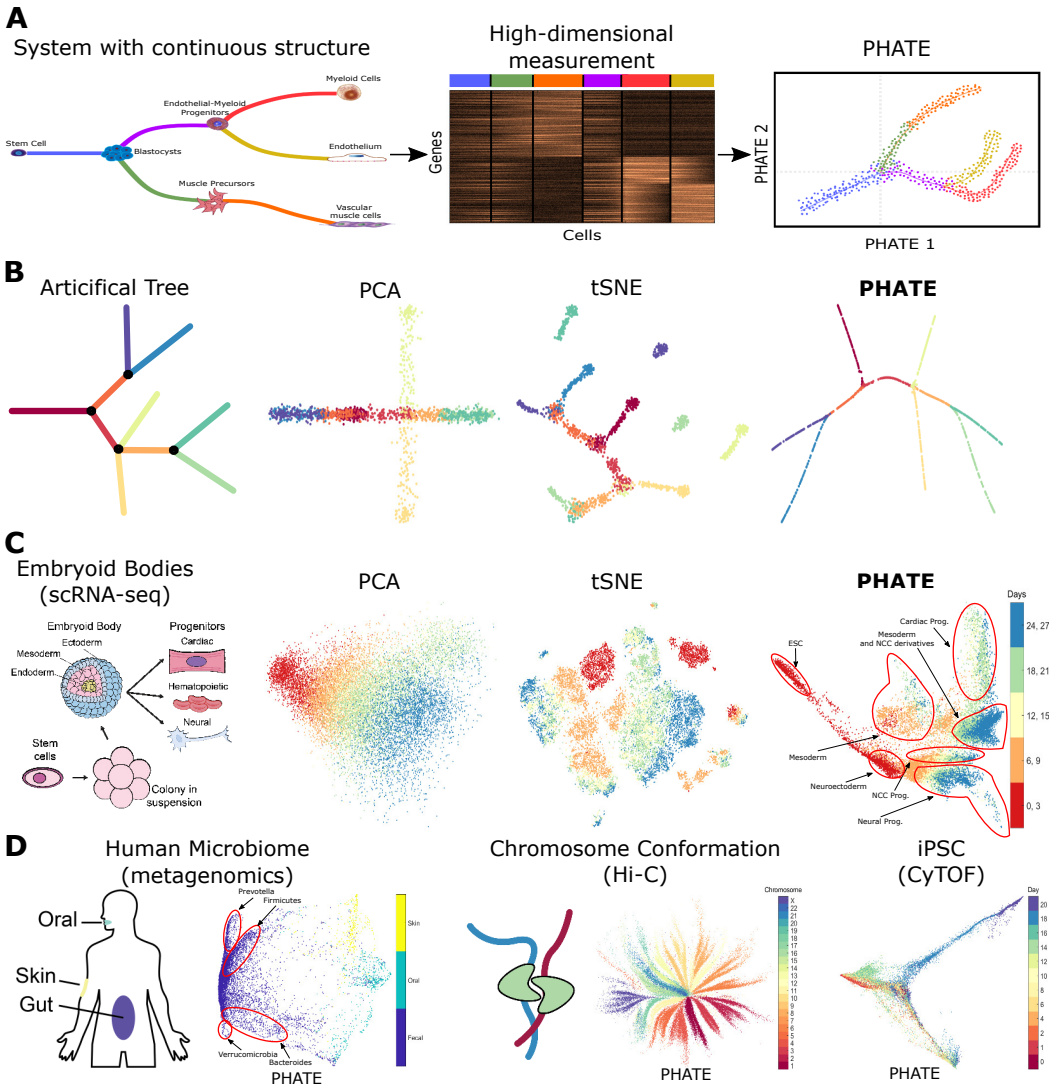

Figure 1: (**A**) Conceptual figure demonstrating the progression of stem cells into different cell types and the corresponding high dimensional single-cell measurements (e.g., mass cytometry or scRNAseq). PHATE embeds the structure within the high dimensional data into lower dimensions which can be used to analyze the progression structure of the data and visualize it in 2 or 3 dimensions. The trajectories and branches can then be analyzed to extract biological meaning. (**B**) (Left) A 2D drawing of a low-noise artificial tree colored by branch. Data is uniformly sampled from each branch in 60 dimensions. See Methods for details on the generated data. (Right) Comparison of PCA, tSNE, and the PHATE visualizations for the high-dimensional artificial tree data. The PHATE embedding is best at revealing the global structure of the data while simultaneously distinguishing the smaller branches from the global structure. In particular, tSNE breaks some of the branches apart (i.e. it destroys the progression structure) and shuffles the broken pieces around within the visualization, thus destroying the global structure. (**C**) Comparison of PCA, tSNE, and the PHATE visualizations for new embryoid body data. PCA captures the overall global progression over time but loses all local structure. Again, tSNE breaks apart the global structure of the data. In contrast, PHATE preserves both local and global structure. For comparisons on other data sets, see Figure 2. (**D**) PHATE applied to various datasets. PHATE is the only method designed to emphasize and preserve continuous structure in data for visualization.

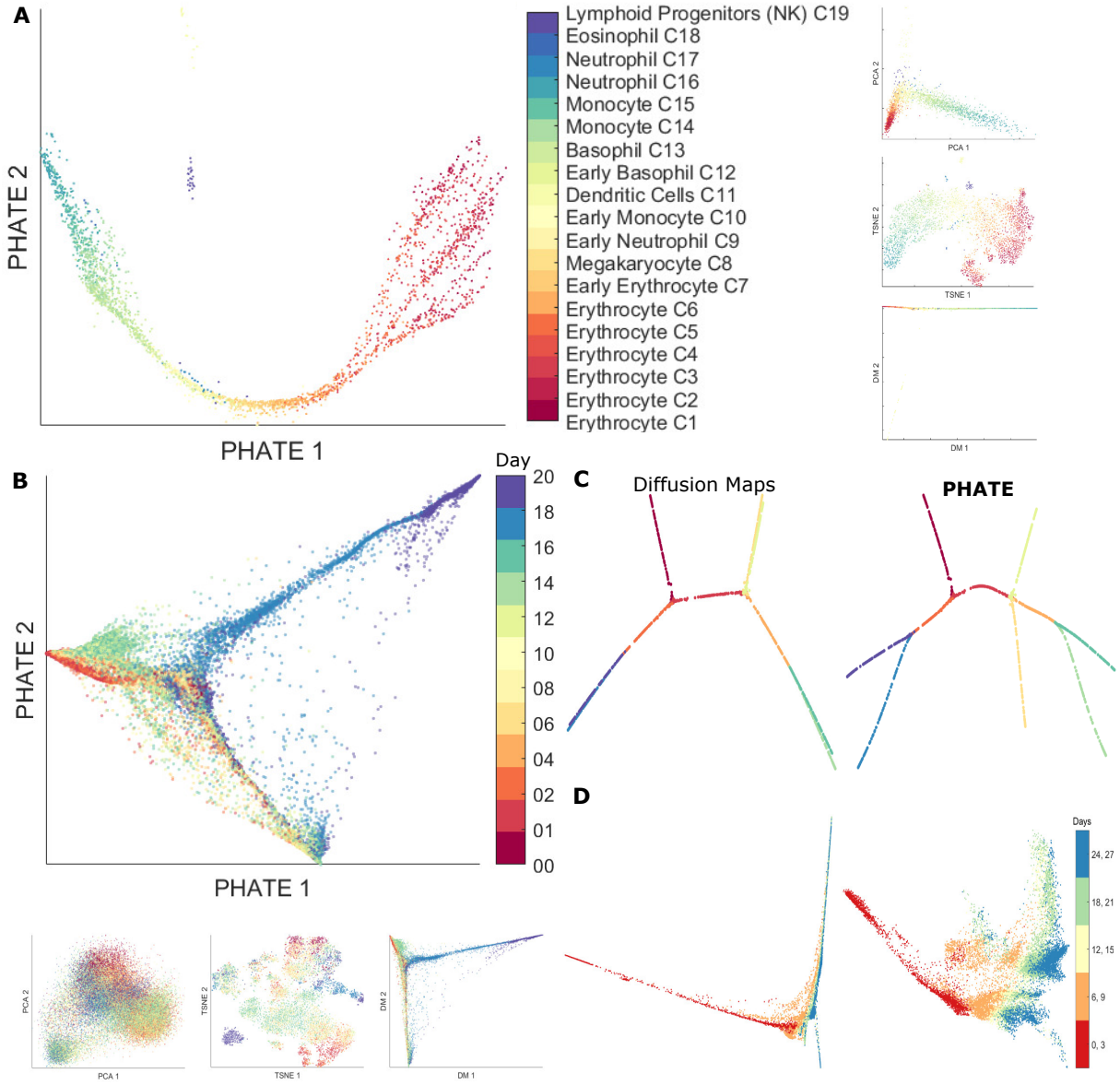

Figure 2: Comparison of PCA, tSNE, DM, and the PHATE visualizations for various datasets. (**A**) Mouse bone marrow scRNAseq data colored by cell type as identified in [6]. The scale for DM and PHATE is $t = 40$. (**B**) iPSC CyTOF data [7] subsampled at $N = 50000$ points and colored by sample time. The scale for DM and PHATE is $t = 250$. Note that tSNE destroys the progression structure in the data. (**C**) Comparison of PHATE to diffusion maps on the artificial tree data from Figure 1B. Note that diffusion maps fails to distinguish several branches in the visualization. Thus diffusion maps fails to visualize all of the local structure. (**D**) Comparison of PHATE to diffusion maps on the EB scRNA-seq data from Figure 1C. Multiple branches are present in the PHATE visualization that are not visible in the diffusion maps visualization.

branch point and branch identification. These methods can then be used to identify genes or other features that correlate with branches to derive meaning from PHATE and further augment its utility in data exploration.

2 The PHATE Algorithm

Current methods of dimensionality reduction are generally not directly designed to enable *visualization* of global nonlinear and local structures (e.g., progression and transitions) in data, and are often ill-suited for this task. In particular, nonlinear embedding methods typically do not enforce visualizable low dimensionality (i.e., two or three dimensions). Furthermore, these methods often introduce distortions by focusing on the separation of local structures in the data, while breaking apart their global structure. In contrast, we propose PHATE as an embedding that is directly designed to enable global and local structure visualization in exploratory settings by satisfying the following four properties:

Visualizable: To enable visualization, the achieved embedding must be sufficiently low dimensional – namely, two- or three-dimensional.

Structure preserving: To provide an interpretable view of dynamics (e.g., pathways or progressions) in the data, the embedding should preserve and *emphasize global nonlinear* transitions in the data, in addition to local transitions.

Denoised: To enable unsupervised data exploration, the embedding should be denoised such that the progressions are immediately identifiable and clearly separated.

Robust: The obtained visualization itself should produce robust features such that their revealed boundaries and the intersections of progressions are insensitive to user configurations of the algorithm.

PHATE first discovers the structure of data by means of data diffusion. Diffusing, or randomly walking, through data enables us to learn the shape of the data via local connections that are propagated and aggregated to reveal global connectivity. To diffuse through data, PHATE first transforms input data measurements (Figure 3, 1st row), distances (Figure 3, 2nd row), or affinities (Figure 3, 3rd row) to create a diffusion operator that captures local neighborhoods in the data. This diffusion operator is a row-stochastic matrix that contains probabilities of transitioning from one data point to another in each single step (e.g., small differentiation step between cells) of a Markovian random walk, which in turn defines a diffusion process over the data. Then, to reveal long-range nonlinear pathways in the data, we propagate the local single-step transitions by powering the diffusion operator (Figure 3, 4th row). The powered operator aggregates together information from multiple random walks over the data. Therefore it captures longer distance connections while also locally denoising the data. However, the information captured by the powered diffusion operator is not directly amenable for embedding in low dimensions for visualization, as we explain later in this section.

To enable such embedding, PHATE computes a novel informational distance metric, which we call potential distance (Figure 3, 5th row), by comparing log-transformed transition probabilities from the powered diffusion operator. The resulting metric space 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 just considering transition probabilities along them. In addition, this potential distance forms an M-divergence that is better adapted to measuring differences between the diffused probabilities [8,9]. We then embed these distances using multidimensional scaling, which preserves distances in the embedding while guaranteeing the low dimensionality of the embedded space. It is worth noting that this metric embedding is enabled by our potential metric, and cannot be reliably obtained from classic diffusion distances (e.g., from [10]), as we demonstrate in Figure 3. The full PHATE algorithm is described in Algorithm 1; computational aspects of each of its steps are explored in Methods.

Algorithm 1: The PHATE algorithm

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

Output: The PHATE embedding Y_m and visualization $Y_{m'}$

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: $D_{U,t} \leftarrow$ compute potential distance matrix from U_t (see Def 1)
9: $Y_m \leftarrow$ apply classical MDS of $D_{U,t}$ to embed in $\mathbb{R}^m$
10: $Y_{m'} \leftarrow$ apply nonmetric or metric MDS to $D_{U,t}$ with Y_m as an initialization for visualization

2.1 Steps of PHATE

Here we motivate and explain how each of the main steps in PHATE helps us ensure the provided embedding satisfies the four properties (namely, visualizable, structure-preserving, denoised, and robust) described above. In Figure 3, we also demonstrate the effects of each step on two examples: the artificial tree data from Figure 1B and data sampled from three noisy intersecting half-circles in three dimensions.

Distances Consider the common approach of linearly embedding the raw data matrix itself, e.g., with Principal Component Analysis (PCA), to preserve the global structure of the data.

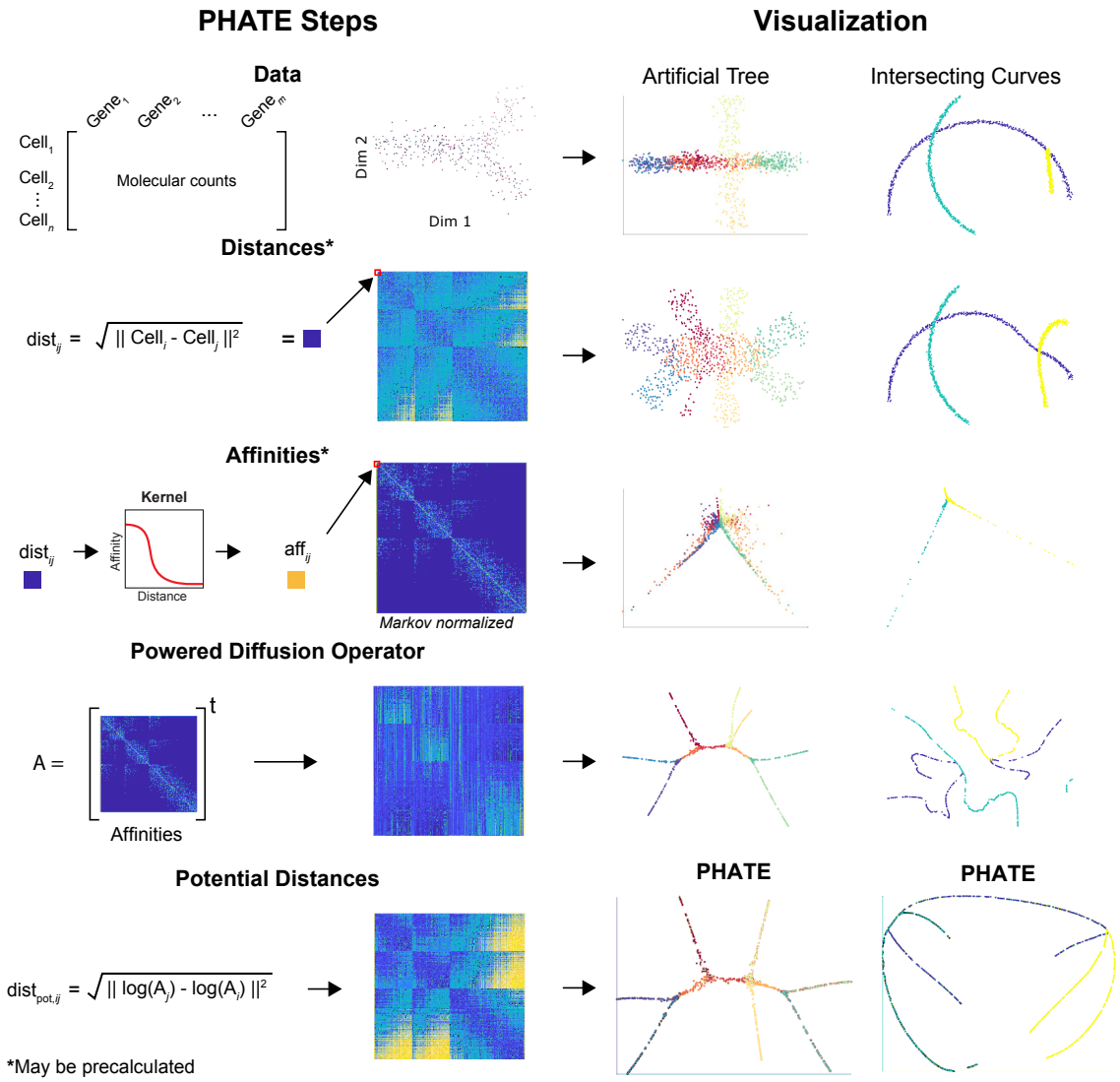

Figure 3: A demonstration of the primary steps in PHATE. Left column: A block diagram showing the steps of the PHATE algorithm when applied to a noisy tree structure with 3 branches in $\mathbb{R}^{12}$. At each step, we apply either metric MDS or PCA to a noisier version of the artificial tree data (middle column) from Figure 1B and three intersecting half circles in 3D (right column). While some of the methods work well for visualizing one of these simple datasets, they break down for the other dataset. PHATE is the only method that accurately visualizes both datasets. Note that if a different distance or affinity is better-suited for a specific dataset, then the distances or affinities can be inputted at the appropriate step in the algorithm.

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. 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 reliable nor 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 changes in marker expression along each development path. Therefore, we transform the global Euclidean distances into local affinities which quantify the similarities between nearby (in the Euclidean space) data points. We do this via a novel, sharply decreasing exponential kernel, which we call the α-decaying kernel, in conjunction with a locally adaptive bandwidth to handle data density variations (see Methods for details).

Propagating Affinities Embedding local affinities directly can result in a loss of global structure as is evident in tSNE (Figures 1 and 2) or kernel PCA embeddings (see the PCA on affinities embeddings in Figure 3, 3rd row). For example, tSNE 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 separated clusters. Therefore, to process the local affinities into global relations, we use them to define small local steps between data points, and then chain them in order 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. By propagating the affinities via the powered diffusion operator, local Euclidean steps are chained to provide a global and robust intrinsic data distance, thus 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 eigenvalues such that the main pathways through the data are emphasized and small noise dimensions are diminished, thus achieving the denoising objective of our method as well. We use this denoising effect to automatically select the diffusion time scale t. Intuitively, it 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 diffusion operator after various powers (see Methods).

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 4B). In fact, this specific property was used in [11] 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.

A naïve approach towards obtaining a truly low dimensional embedding of diffusion geometries is to directly apply a metric embedding, such as MDS, from the diffusion map space to a two dimensional space. However, as seen in Figure 3 (4th row), direct embedding of this distance is unstable, and does not provide faithful or robust visualizations.

Embedding Potential Distances To resolve instabilities in diffusion distances and embed the global structure captured by the diffusion geometry in visualizable low dimensions, we replace it with a novel informational diffusion-based distance, which we call potential distance. The potential distance is inspired by two sources that regularly model and compare systems via probability distributions. 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 in sparse areas. Secondly, when analyzing dynamical systems of moving particles, it is not the pointwise 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 diffusion operator and consider the L^2 distance between these localized energy potentials in the data as our manifold-distance, which forms an M-divergence between the diffusion probability distributions [8, 9]. The final step in the PHATE embedding is to then embed the potential distances in lower dimensions via metric or nonmetric MDS (see Figure 3, 5th row). We see that embedding the potential distances has better boundary conditions near end points compared to diffusion maps, even in the case of simple curves that contain no branching points. Figure 4 shows a half circle embedding with diffusion distances vs 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. Thus the potential transform succeeds in visualizing data where both diffusion maps succeeds and fails.

PHATE achieves an embedding that satisfies all four properties delineated above: PHATE preserves and emphasizes the global and local structure of the data via a localized affinity that

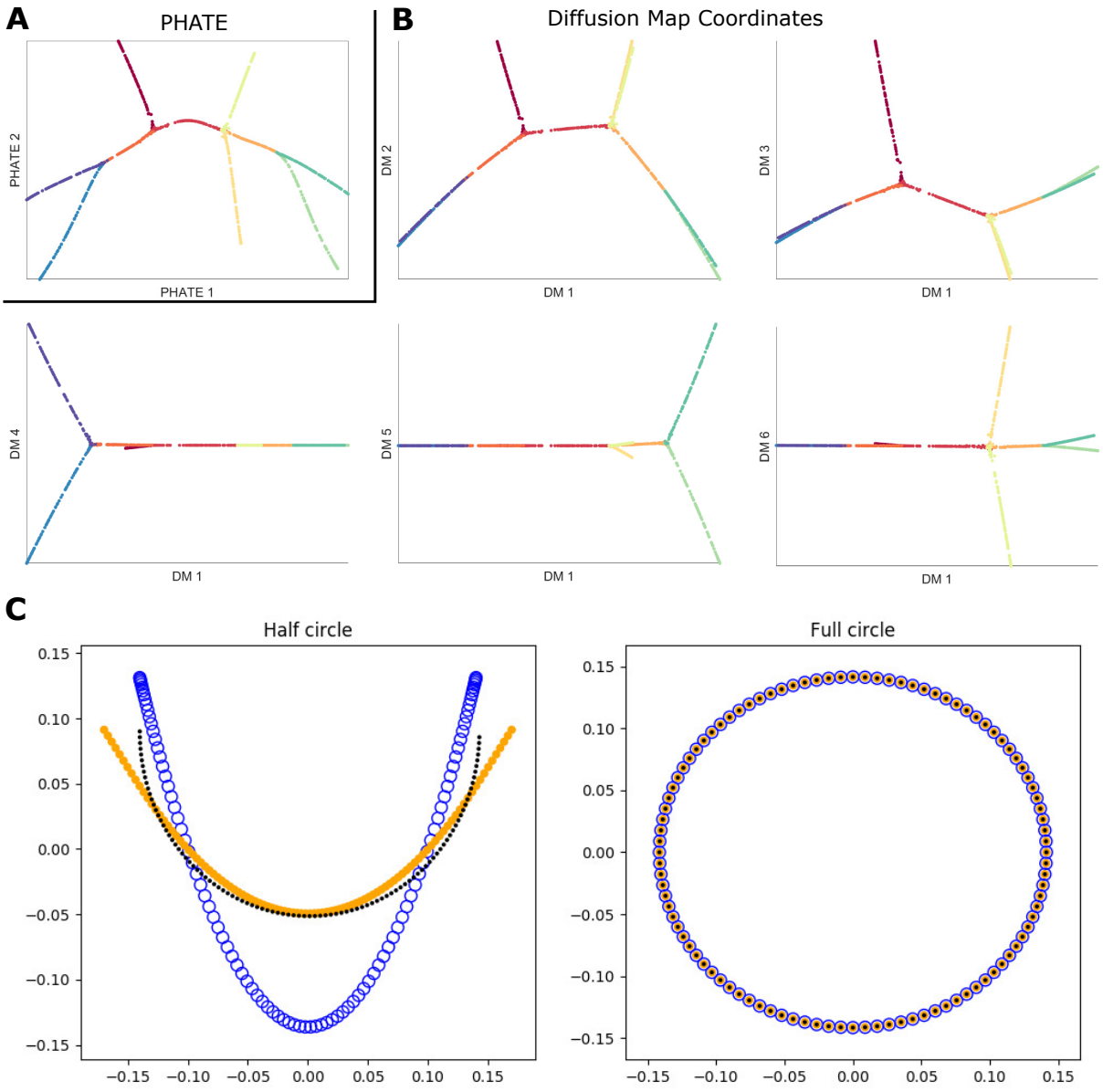

Figure 4: Diffusion maps visualization weaknesses. (**A**) PHATE applied to the artificial tree data. Only two PHATE coordinates are needed to separate all branches. (**B**) The first six diffusion map coordinates of the artificial tree data. At least five of these coordinates are necessary to separate 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.

is chained via diffusion to form global affinities through the data manifold, denoises the data by low-pass filtering through diffusion, provides a distance that has robust boundary conditions for purposes of visualization by taking the difference between distribution potentials of the diffusion operator, and captures the data in low dimensions using MDS for visualization. We have shown by demonstration in Figure 3 that all of the steps in the PHATE algorithm are necessary to achieve these properties. In particular, we will show further via comparisons that the potential transform is necessary as diffusion maps will fail to provide an adequate visualization, even when using the customized α-decaying kernel we developed for PHATE.

2.2 Extracting Information from PHATE

PHATE embeddings contain a lot of information about the structure of the data: namely local transitions, progressions, and branches or splits in progressions, and end states of progression. In this section, we describe 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. Branch points often encapsulate switch-like decisions where cells sharply veer towards one of a small number of fates. For example, Figure 6A shows the PHATE visualization of the iPSC CyTOF datset from [7] 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 6B) and marks a major decision point between these two cell fates. Identifying branch points in biological data is of critical importance for analyzing these decisions.

Many of the main branch points in a dataset can be identified by visual inspection of the PHATE visualization. However, there may exist some branch points (i.e. that involve relatively fewer cells) that may be more easily identified in higher PHATE dimensions. Here we present a method provides suggested regions for these branch points.

We make a key observation that most points in PHATE plots of biological data lie on roughly one-dimensional progressions with some noise as demonstrated in Figure 5Aii. 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 5Aii) 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

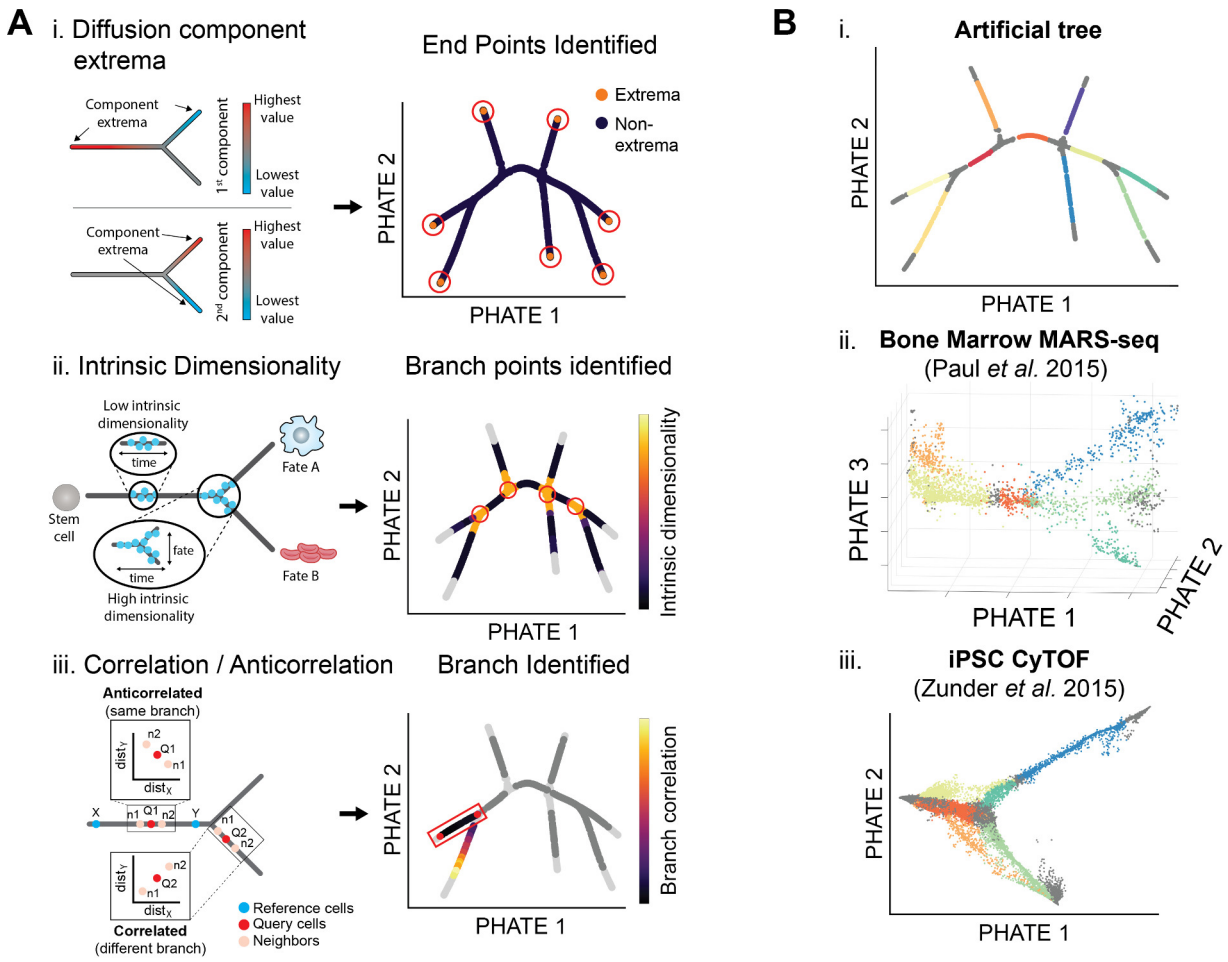

Figure 5: (**A**) Cartoons and examples demonstrating our methods for identifying suggested endpoints, branch points, and branches. (i) A cartoon highlighting the extrema of different diffusion components which are typically located at endpoints [12]. The diffusion component extrema are used to identify the endpoints of the artificial tree data. (ii) A cartoon showing local intrinsic dimension on a branch and at a branch point. There are more degrees of freedom at a branch point, resulting in larger intrinsic dimension. The local intrinsic dimension highlights branch points in the artificial tree data. (iii) A cartoon showing anticorrelated distances when considering a point (Q_1) on the branch between x and y and correlated distances when considering a point (Q_2) on a separate branch. This property is highlighted on the artificial tree data when considering the branch in the box. (**B**) Detected branches in the (i) artificial tree data, (ii) bone marrow scRNA-seq data from [6], and (iii) iPSC CyTOF data from [7]. Reasonable branches are identified in each case.

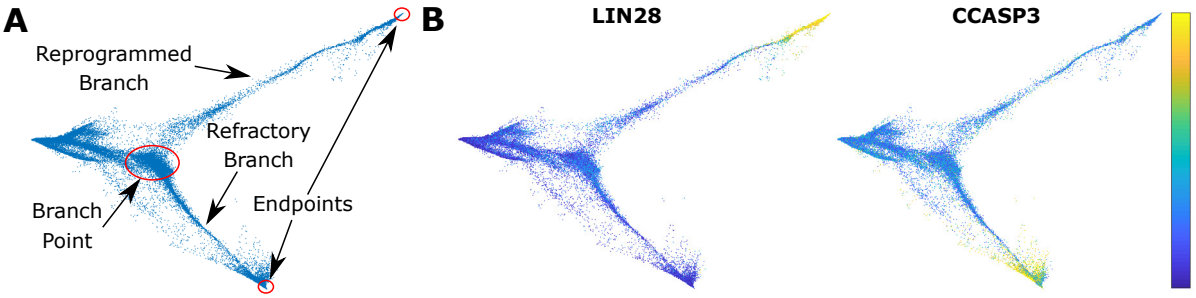

Figure 6: Annotated PHATE visualizations of the iPSC CyTOF dataset [7]. (**A**) The primary branch point between the two major branches (reprogrammed and refractory) of the data is highlighted. (**B**) The PHATE visualization colored by LIN28 (a marker associated with the transition to pluripotency [13]) 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.

will require a higher dimensional structure such as a plane, as shown in Figure 5Aii. 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 [14]. 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 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 of this approach, see Methods. Figure 5Aii 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 end points in the PHATE embedding. These points can correspond to the beginning or end-states of differentiation processes. For example, Figure 6A shows the PHATE visualization of the iPSC CyTOF datset from [7] 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 two criteria: data point centrality and distinctness.

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 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 that the beginning or end-states of differentiation processes to have the most distinctive

cellular profiles. As shown in [12] we quantify this distinctness by considering the extrema of diffusion eigenvectors (see Figure 5Ai). 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 [15]. Essentially, this approach groups collections of branch points or endpoints together into representative points based on their proximity (see Methods for details on this procedure). 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 the method described in [11] that compares the correlation and anticorrelation of neighborhood distances. However, we use higher dimensional PHATE coordinates instead of the diffusion maps coordinates. Figure 5B shows the results of our approach to identifying branch points, endpoints, and branches on the artificial tree data, the bone marrow scRNA-seq data from [6], and the iPSC CyTOF data from [7]. Our procedure identifies the branches on the artificial tree perfectly and defines reasonable branches on the other two datasets which we will use for data exploration.

2.3 PHATE Reveals the True Structure of Data

We saw in the previous section that the globally propagated local affinities used by PHATE learn the manifold structure and that log transforming these affinities and then embedding the resultant potential distances with MMDS preserves these distances in a stable manner for visualization purposes. Here, we show that PHATE is indeed able to correctly visualize both local and global structure on data that has 1) continuous tree-like progression structure, 2) cluster structure, 3) and a mixture of the two types of structure. In other words, PHATE does not impose any type of structure on the data but rather reveals the underlying structure of the data visually.

Figure 7A shows that PHATE renders an artificial tree dataset (see Methods) correctly. This dataset contains a tree with 10 branches, which each branch containing around 100 points in different data subdimensions of a 60 dimensional space. PHATE recovers this structure faithfully, with branches correctly oriented, while tSNE shatters the space and PCA smears the branches such that they artificially overlap (Figure 7A). Thus PHATE discovers both the global and local structure correctly including intersections of branches in the data.

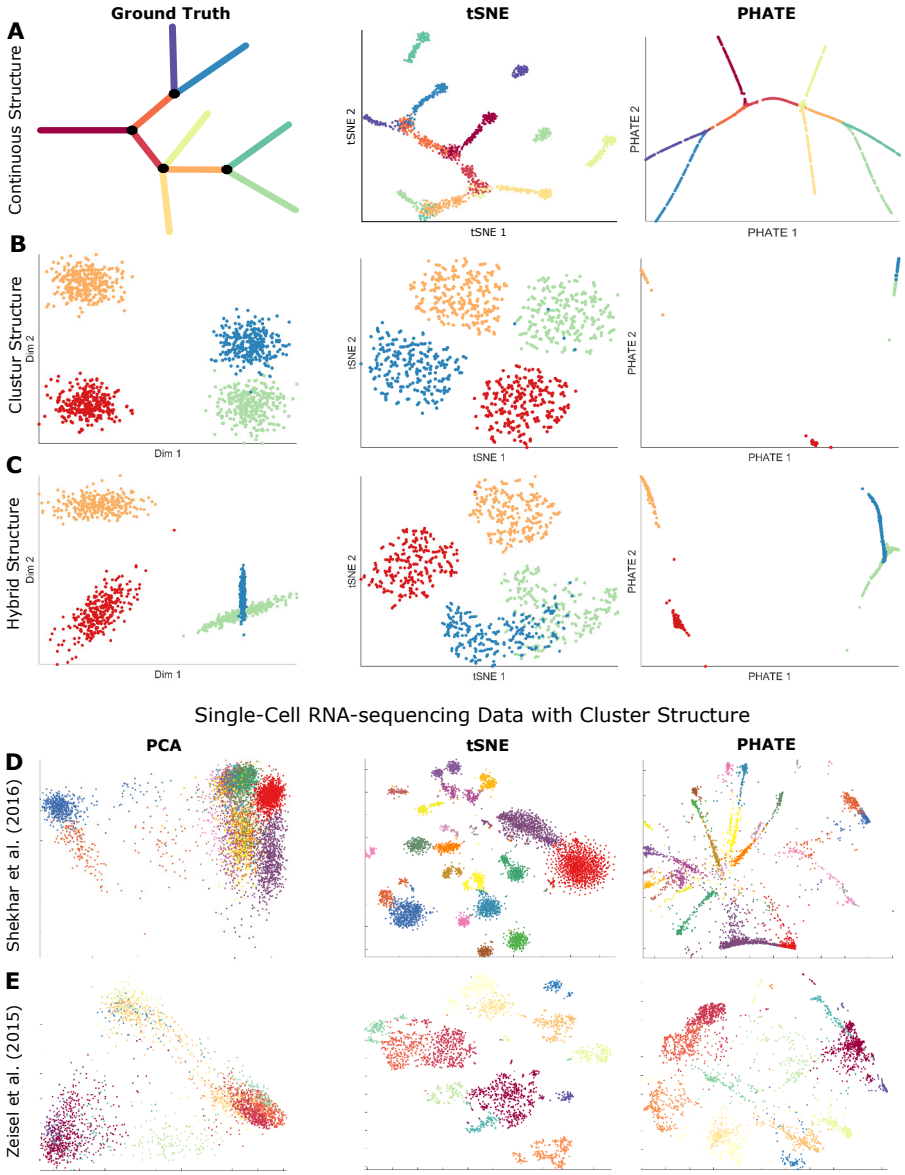

Figure 7: PHATE reveals the true structure of the data under different settings. (**A**) PHATE accurately reconstructs the artificial tree data, demonstrating that PHATE can reconstruct continuous structured data characterized by branches and progressions. (**B**) PHATE accurately reconstructs naturally clustered data and also preserves the relative location of clusters. In contrast, tSNE distorts the relative position of clusters including. This is demonstrated on samples from four 24-dimensional Gaussian distributions with identical covariance matrices and different means. Only the first two dimensions are shown in the ground truth. (**C**) PHATE accurately reconstructs and denoises hybrid structured data with a mixture of branching and cluster structure. The ordering within clusters is also preserved whereas tSNE distorts this ordering and does not denoise the data. This is demonstrated on samples from 24-dimensional Gaussian distributions with different means and covariance matrices. Only the first two dimensions are shown. (**D**) Comparison of PCA, tSNE, and PHATE on scRNA-seq data obtained from retinal bipolar neurons [16]. (**E**) Comparison of PCA, tSNE, and PHATE on scRNA-seq data obtained from cells in the mouse cortex and hippocampus [17]. Points are colored by cluster assignments obtained using Phenograph [18].

In Figure 7B, we see data sampled from four 24-dimensional Gaussian distributions with identical covariance matrices and different means, two of which are overlapping. PHATE correctly renders the data as three separate groups with the two overlapping Gausians forming a continuum that is joined end-to-end rather than overlapping. While tSNE is able to recover these clusters as separate entities, it is unable to place the clusters correctly relative to one another. This is due to the fact that tSNE preserves only local distance and does not propagate such distances globally. Furthermore, the KL-divergence penalty used in the tSNE optimization is nearly zero for far points that are placed close together while the penalty is high for very close points (neighbors) that are placed far way from each other on average. Thus tSNE does not "care" about placing globally distant objects, like clusters, correctly in global terms. Thus the blue and green clusters are farther from one another than the blue and red or blue and orange clusters.

Figure 7C shows data that has divergent progression within the clusters. Here, while tSNE roughly approximates this data, we see that only PHATE shows the branch point and the point of intersections clearly. Additionally, the progression within the red and orange clusters is shown much more clearly by PHATE than tSNE.

Next we compare PHATE with PCA and tSNE on two neuronal single-cell RNA-sequencing datasets with cluster structure that have been reported in literature. Figure 7D shows the results on retinal bipolar data [16] while Figure 7E shows data from a mouse cortex and hippocampus [17]. Both datasets are colored by labels after clustering with Phenograph [18]. Here we see that the global structure rendered by PHATE is similar to PCA (which captures global variation), but local structure is clarified by PHATE. Additionally, cluster separations are maintained in the PHATE visualization as in tSNE. Thus, PHATE captures both global structure and refines and denoises local structure.

2.4 Comparing PHATE to Other Methods

We compare PHATE to methods of dimensionality reduction, tree rendering, and visualization on several datasets in Figures 1, 2, and 8. The datasets used in the comparisons include: 1. Artificial tree data; 2. Developing mouse bone marrow cells, enriched for the myeloid and erythroid lineages, which were measured with the MARS-seq single cell RNA-sequencing technology [6]; 3. Mass cytometry data showing iPSC reprogramming of mouse embryonic fibroblasts [7]; 4. New embryoid body data from a 27-day timecourse. The biological datasets represent differentiating processes within the body, and hence visualizing progression is key to understanding the structure of these datasets.

PHATE is primarily a dimensionality reduction method that takes high dimensional raw data and embeds it, via a metric preserving embedding, into low dimensions that naturally show trajectory structure. Thus, we focus our comparisons of PHATE to existing dimensionality reduction methods such as PCA, tSNE, diffusion maps, and single cell topological data analysis (scTDA) [19]. 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 [20, 21] and Monocle2 [3].

Finally, we note that several methods exist that focus on finding *pseudotime* orderings of cells, such as Wanderlust [5], Wishbone [4], and diffusion pseudotime [11]. 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 using the previously described procedure.

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.

2.4.1 Comparison of PHATE to dimensionality reduction methods

Figures 1 and 2 compare the PHATE visualization to the principal components analysis (PCA), tSNE, and diffusion maps on four different data sets while Fig. 8 compares PHATE to scTDA on three datasets. 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 [22]. However, PCA assumes a linear structure on the data, the visualization amounts to projecting the data on to 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 1B where we compare PCA to PHATE on artificial tree data. This data contains 10 distinct branches uniformly sampled in 60 dimensions. See Methods 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 is noisy compared to the PHATE visualization, in which all 10 branches are easily identifiable.

For the other datasets in Figures 1C and 2, PCA captures some of the overall global structure of the datasets. For example, the PCA dimensions in Figure 1C 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. Simi-

lar results are obtained from the mouse bone marrow scRNA-seq and iPSC CyTOF datasets in Figures 2A and 2B, respectively, demonstrating that PCA is unable to accurately visualize the global and local structure of the data simultaneously.

Comparison of PHATE to tSNE: tSNE (t-distributed stochastic neighbor embedding) [1] is a visualization method that emphasizes local neighborhood structure within data. Recently, tSNE has become popular for revealing cluster structure or separations in single cell data [23]. However, due to its emphasis on preserving local neighborhoods, tSNE tends to shatter trajectories into clusters as seen in Figures 1B, 1C, and 2B. In all of these cases, the data naturally have a strong trajectory structure either by design (the artificial tree data) or due to the developmental nature of the data (the EB and iPSC datasets). Thus tSNE creates the false impression that the data contain natural clusters which could lead to incorrect analysis.

Furthermore, the adaptive kernel used in tSNE for calculating neighborhood probabilities tends to spread out neighbors such that dense clusters occupy proportionally more space in visualization as compared to sparse clusters [24]. Thus, the relative location of data points within the tSNE embedding often does not accurately reflect the relationships between them. This is clearly visible in the tSNE plot in Figure 1B where the shattered branches are located far away from where they originated in the main structure in the artificial tree data. Similarly, tSNE creates clusters in the EB and iPSC data in Figures 1C and 2B which split the time samples into different components. Since the relative position of clusters in tSNE is generally meaningless the overall progression of the data is destroyed.

Even in the case where the data are more naturally separated into clusters, tSNE 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 7) 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 tSNE only preserves local structure.

Comparison of PHATE to Diffusion Maps: Diffusion maps effectively encode continuous relationships between cells. However, different trajectories are often encoded in different dimensions as seen in Figure 4B, which is unsuitable for visualization. In contrast, PHATE effectively encodes trajectories in lower dimensions for visualization. This is also seen clearly in Figure 2C in the comparison of PHATE to diffusion maps on the artificial tree 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 10 branches (see Figure 4B). In contrast, all of the 10 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 2A and 2D, respectively). Additionally, the diffusion maps instabilities mentioned previously appear to cause very noisy data (e.g. the scRNA-seq data in Figures 2A and 2D) 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 due to its instabilities and its propensity to encode different trajectories in different dimensions.

Comparison of PHATE to scTDA: scTDA is a scRNA-seq toolkit based on the network learning algorithm Mapper [25]. Mapper can use any dimensionality reduction tool to obtain a two-dimensional projection of the data, aligns overlapping bins across the projection, and then clusters cells within each bin in the high-dimensional space. Clusters are then linked based on the number of shared cells between clusters. The resulting graph is then used for gene expression analysis and can also be used for visualization.

We applied scTDA to the artificial tree data, the bone marrow scRNA-seq data, and the new EB scRNA-seq data in Figure 8. Figure 8A shows scTDA applied to the artificial tree data. Compared to PHATE, the branching structure of the data is much less clear in the scTDA visualization. Additionally, several branches appear to be merged and some spurious connections are made between branches. scTDA also can give different results when running the algorithm multiple times on the same data with the same settings, as evidenced in Figures 8B and 8C. Thus it is difficult to determine the right structure of the data. Additionally, the overall structure of the data is much more difficult to infer from the scTDA visualization compared to PHATE.

scTDA is also sensitive to the parameters. Figure 8D shows the scTDA visualization for the bone marrow and EB scRNA-seq datasets using different parameters from those in Figures 8B and 8C. The results show significant variability. In contrast, PHATE is robust to the choice of parameters (see Methods).

2.4.2 Comparison of PHATE to tree-rendering methods

SPADE [20, 21] and Monocle2 [3] 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 [20]. In Anchang et al. [21], 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

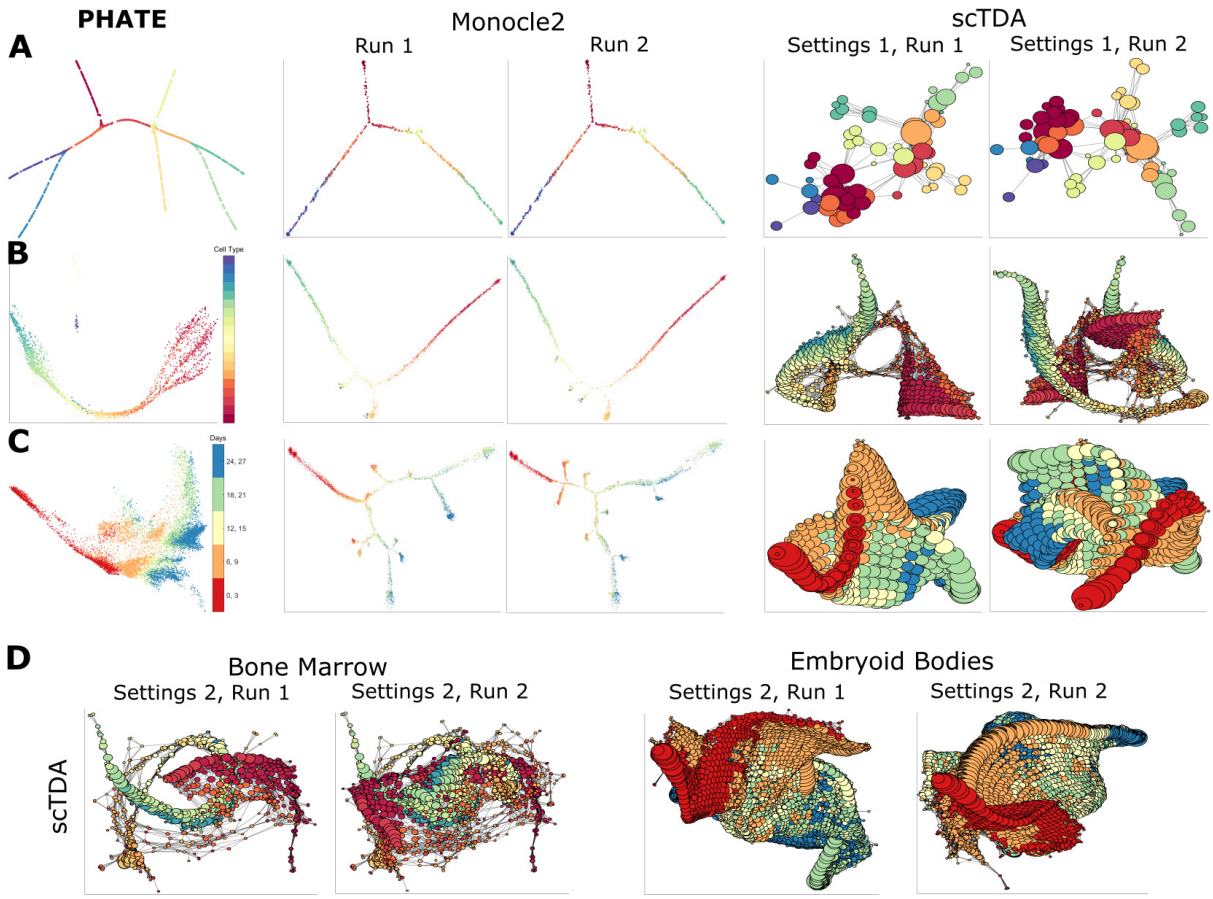

Figure 8: Comparison of PHATE to Monocle2 and scTDA on the (**A**) artificial tree data colored by branch, (**B**) the bone marrow scRNA-seq dataset from [6] colored by cell type, and (**C**) the new scRNA-seq EB data. Multiple runs are shown of both Monocle2 and scTDA. For the bone marrow and EB datasets, both Monocle2 and scTDA give different results under the same settings, suggesting that they are sensitive to randomization. In contrast, given the same parameters, PHATE produces the same results with each run. (**D**) scTDA results on the bone marrow scRNA-seq and EB datasets using different parameter settings. The results show significant variability, suggesting that scTDA is sensitive to parameter selection. In contrast, PHATE is robust to the choice of parameters (see Methods).

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. [21]. 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% [21]. 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 [26, 27] as a default. We compare Monocle2 to PHATE in Figure 8 on the artificial tree data, the bone marrow scRNA-seq data, and the new EB scRNA-seq data. For the artificial tree data, Monocle2 fails to detect multiple branches (Figure 8A). For the bone marrow scRNA-seq data, 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.

For the EB scRNA-seq data, Monocle2 detects multiple branches. However, as before, the number and location of these branches 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.

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, which is able to reveal paths of progression, decision or branch points, and end-states within the biological datasets used in Figure 2. We also use PHATE to reveal structure in SNP data, microbiome data, Facebook network data, Hi-C chromatin contact data, and facial images.

3.1 PHATE on Single-Cell Data

We show that the identifiable trajectories in the PHATE visualization have biological meaning that can be discerned from the expression and mutual information of genes along the trajectories. Figures 9 and 10 show the results for the bone marrow scRNA-seq [6] and iPSC CyTOF [7] datasets. For both of these datasets, we used the method described previously for detecting branches. We then ordered the cells within each trajectory using Wanderlust [5]. Ordering is generally from left to right. We note that we could also order these points based on other pseudotime ordering software such as those in [4] or [11].

Figure 9 shows the PHATE embedding for the two datasets with the trajectories identified

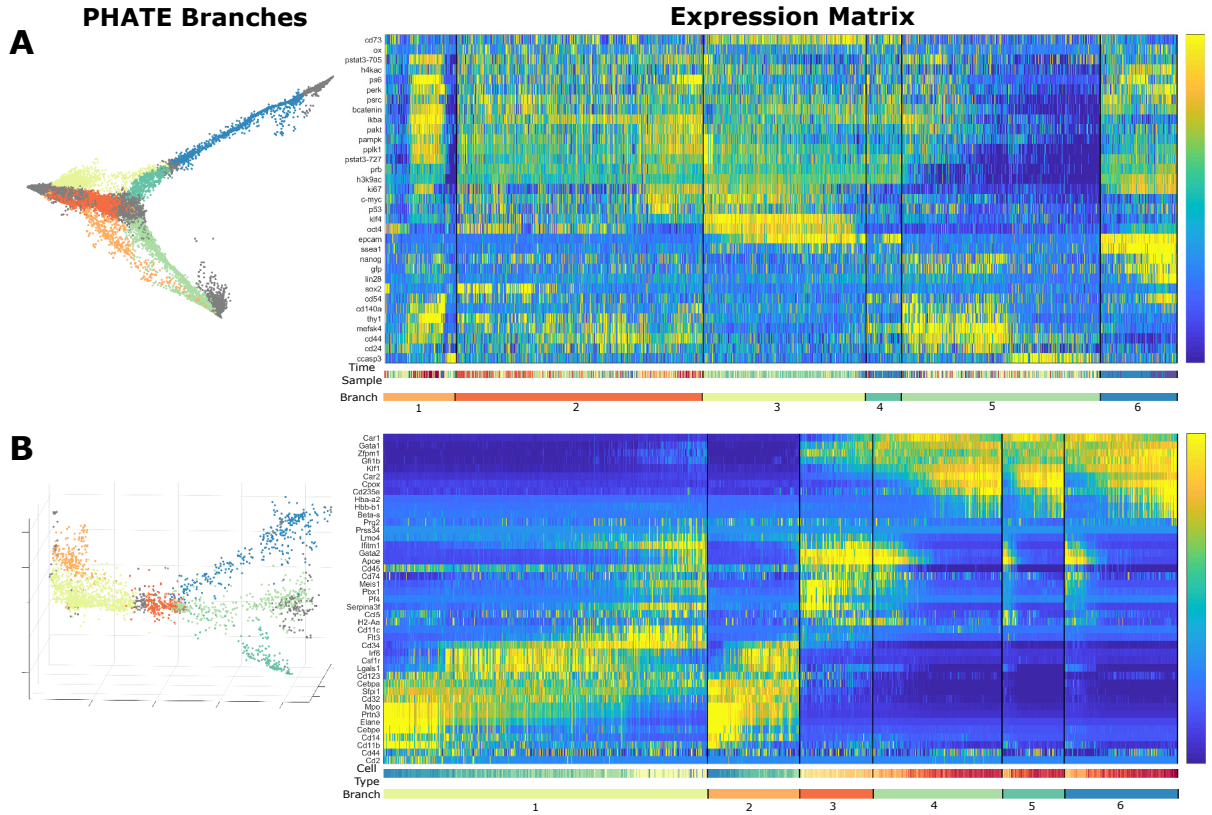

Figure 9: Analysis of branches on the PHATE embedding for the (**A**) iPSC mass cytometry dataset from [7] and (**B**) bone marrow scRNA-seq dataset from [6]. (Left) The PHATE visualization with identified branches. (Right) Expression level for each cell ordered by branch and ordering within the branch. Cell ordering is calculated using Wanderlust [5] 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. MAGIC [28] is applied first to the scRNA-seq data in (**A**) to impute missing values in the data using the same kernel used for PHATE and scale $t = 4$. For branch analysis, we expanded the visualization dimension of the data in (**B**) to 3.

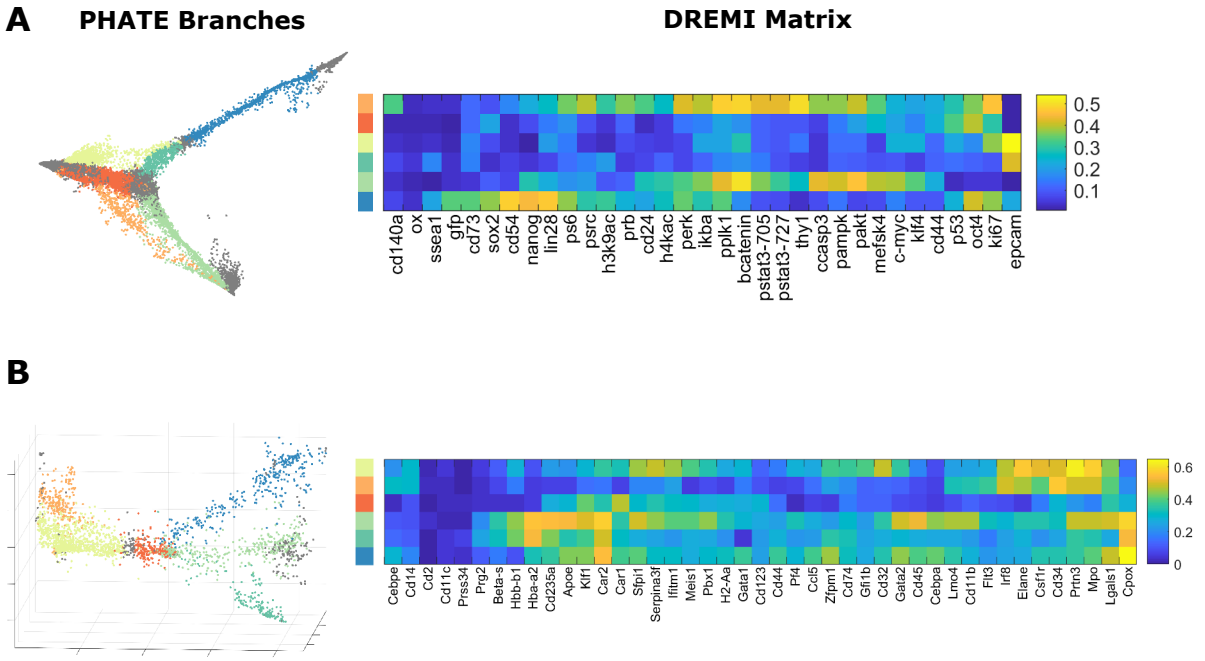

Figure 10: DREMI scores [29] between gene expression levels and cell order within each branch for the same 2 datasets as in Figure 9. MAGIC is applied first to the scRNA-seq data in (**A**) to impute missing gene values.

by color along with gene expression matrices that show the expression level of each cell along the trajectory. Ubiquitously expressed genes along a trajectory can allow us to identify cells of the trajectory. Additionally, we show DREMI (a conditional-density resampled mutual information that eliminates sampling biases to reveal shape-agnostic relationships between two variables [29]) matrices in Figure 10 that show the mutual information between the cell order within each branch and selected markers to show which genes change along the branch to form the progression.

iPSC Mass Cytometry Data Figure 9A shows a mass cytometry dataset from [7] 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 [7]. 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. Out of branch 1 emerges two branches. Branches 4 and 6 on top show the successfully reprogramming ESC-like lineages expressing markers such as Nanog, Oct4, Lin28 and Ssea1, and Epcam that are associated with transition to pluripotency [13]. Branch 5 shows a lineage that is refractory to reprogramming, does not express pluripotency markers and is referred to as still "mesoderm-

like" in [7].

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 as Epcam, which is known to promote reprogramming generally [30], increases along this branch (as evidenced by its high DREMI score in Figure 10C against the branch). It joins into the 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 successfully 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 [7].

Bone Marrow scRNA-seq Data reveals new structure Figure 9B shows the color-coded 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 [6], which are provided in Figure 2A. Here, we show that PHATE reveals a continuous progression structure instead of cluster structure and illustrates the connections between clusters. The PHATE embedding shows a continuous progression from progenitor cell types (shown in light green in the "Cell Type" color bar below the expression matrix) to erythroid lineages (in red) towards the right and myeloid lineages towards the left (in cooler colors). The expression matrix shows increasing expression of erythroid markers in the rightmost branches (branches 4, 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 [6]. 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 also splits 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 Fig. 7). However, we note that both the tSNE and PCA embeddings of this data in Fig. 2 also lack a gap between these trajectories. Given that tSNE in particular is designed to separate clusters, this lack of separation is likely due to low cell number and depth of measurements in the data.

3.2 PHATE on High-dimensional High-throughput Data

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

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 [42]. Here, we see that as compared to PCA, PHATE embedding shows clear population structures, such as the near eastern Jewish populations near the bottom (Iranian and Iraqi Jews, Jordanians), with further branches showing progression within the same population, such as the Jordanian population in orange diamond. Further, PHATE shows a global structure that mimics geography, with European populations generally towards the top and Near Eastern populations towards the bottom. Thus PHATE shows that the occurrence and structure of these SNPs follows a progression based on geography and population divergence. PCA tends to crowd populations together into two linear branches, without clearly distinguishing between population groups or showing population divergence.

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 (Fig. 11C 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 12B shows the PHATE embedding colored by two genera (bacteroides and prevotella) and a phylum (actinobacteria) of bacteria on the same 9660 samples as in Fig. 12A. 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

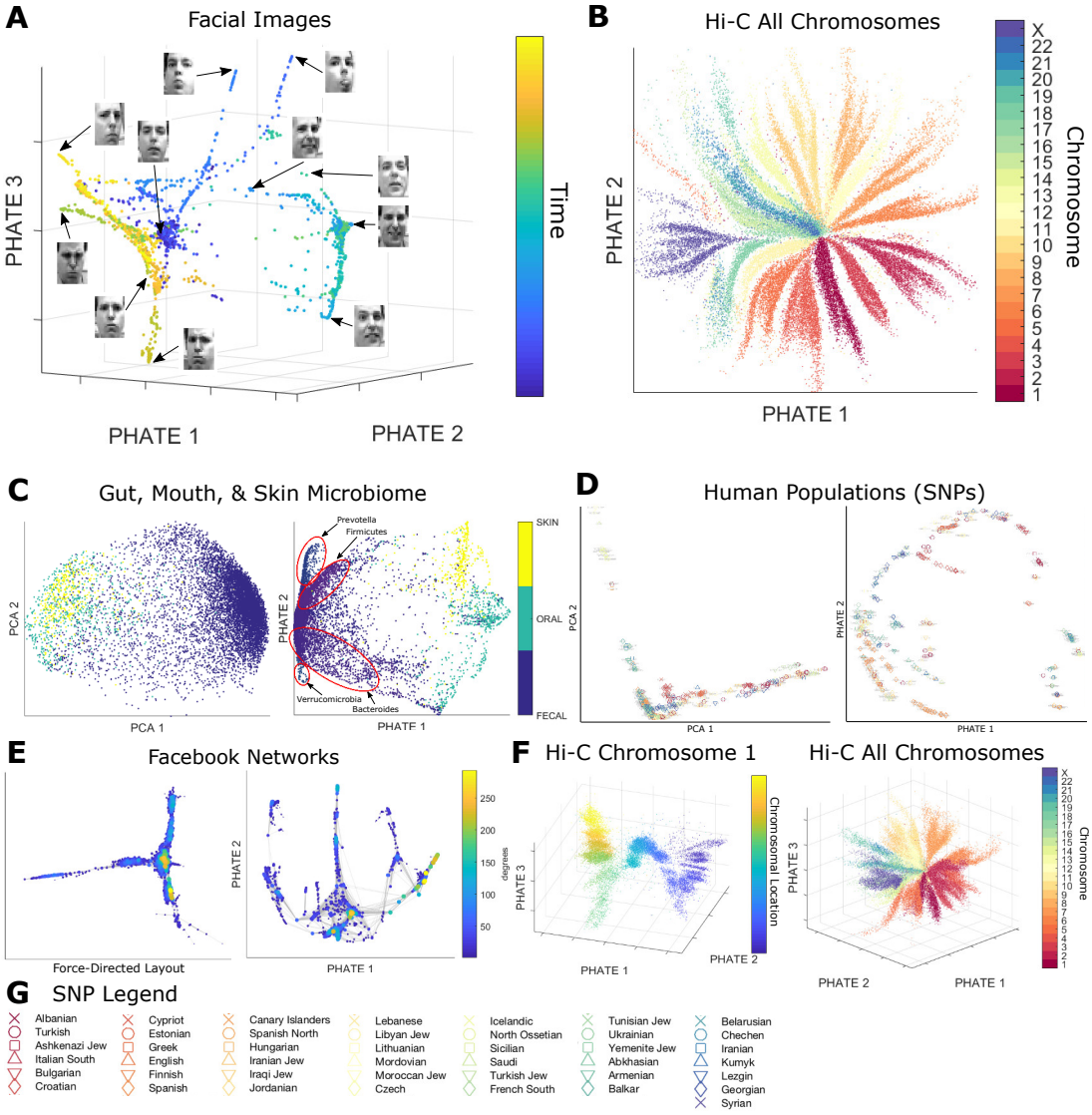

Figure 11: (**A**) A 3D PHATE visualization of the Frey Face dataset used in [40]. Points are colored by time within the video. Multiple branches corresponding to different poses are clearly visible. (**B**) 2D PHATE visualization of human Hi-C data [41] using all 23 chromosomes at 50 kb resolution, colored by chromosome. Each point corresponds to a genomic fragment. (**C**) PCA and PHATE embeddings of data from the American Gut project, colored by body site, and branches annotated by dominant bacteria genera or phyla (Fig. 12). (**D**) PCA and PHATE embeddings of the Human Origins dataset showing genotyped present-day humans from 203 populations [42]. (**E**) Force-directed layout and PHATE visualizations of Facebook network data with data points colored by their degree (number of connections). (**F**) 3D PHATE visualizations of the same human Hi-C data in **B** for chromosome 1 (left) at 10 kb resolution and all 23 chromosomes (right) at 50 kb resolution. (**G**) Population legend for the SNP data in **D**.

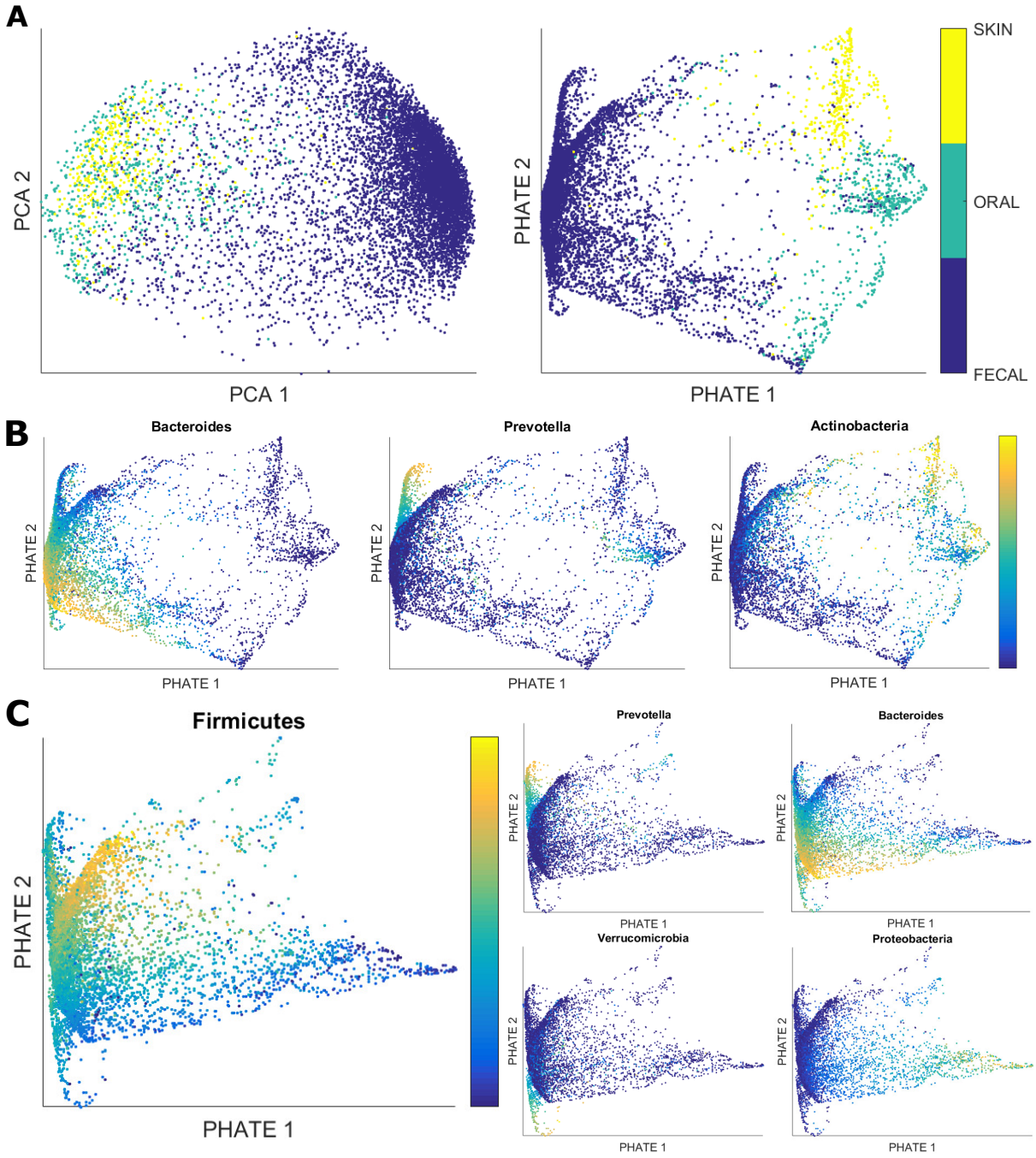

Figure 12: Analysis of data from the American Gut project. (**A**) PCA and PHATE embeddings colored by body site. PHATE shows multiple branches that are not visible in the PCA embedding. (**B**) The PHATE embedding colored by 2 genera (bacteroides and prevotella) and a phylum (actinobacteria) of bacteria. (**C**) The PHATE embedding of only the fecal samples colored by various genera (bacteroides and prevotella) and phyla (firmicutes, verrucomicrobia, and proteobacteria) of bacteria. Each PHATE branch is associated with one of these bacteria groups.

with the work in [43] which showed that different genera and phyla of bacteria are prevalent in the different body sites.

Upon "zooming in" to the 8596 fecal samples in Fig. 12C, we see 4 major branches, instead of the three enterotypes reported in previous literature [44], 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 [45, 46]. 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 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 [40]. This dataset consists of nearly 2000 video frames of a single subject's face in various poses. Figure 11A 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.

3.3 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 low dimensional 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 which 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 [41] by using the Hi-C contact map as the affinity matrix in the PHATE

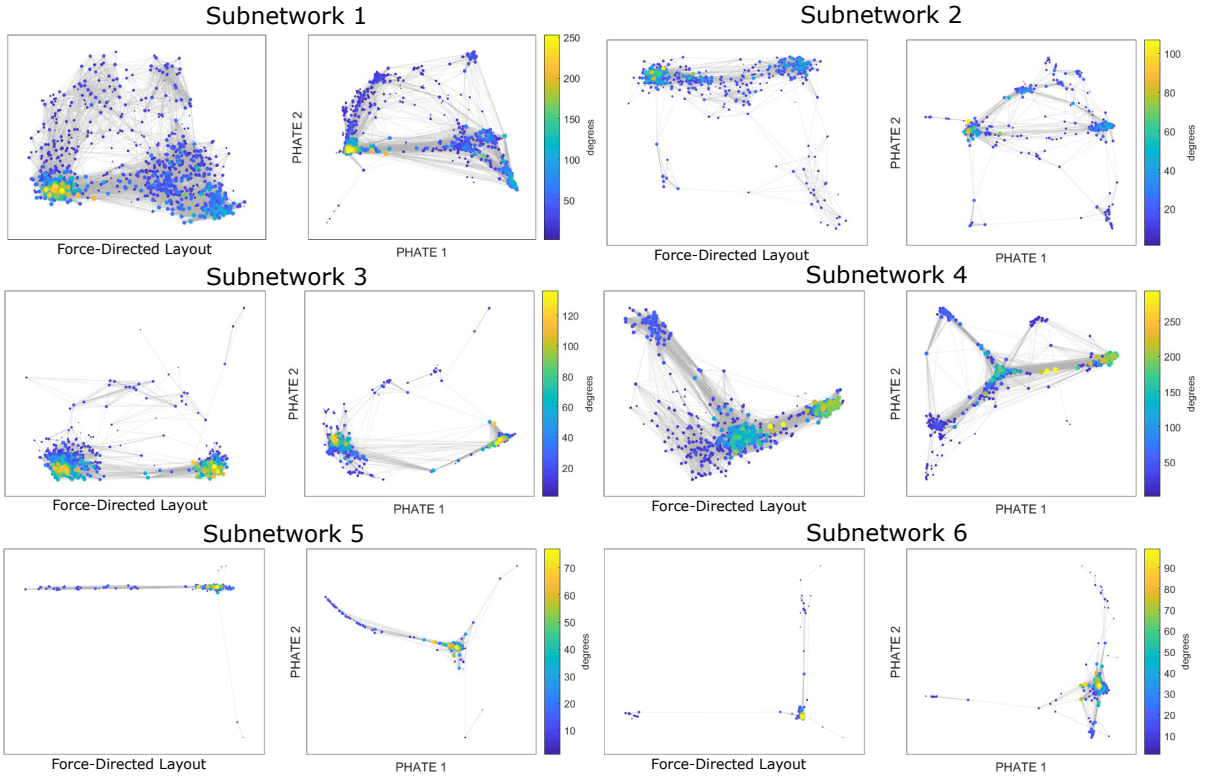

Figure 13: Comparison of the force-directed layout and PHATE visualizations of subnetworks within the Facebook network data in Fig. 11E. The subnetworks are taken from the friend networks of selected individuals within the entire network. In all cases, PHATE reveals more structure.

algorithm. The contact map gives the (relative) number of times two genomic locations are observed in spatial proximity. Hi-C contact maps are usually visualized using the matrix directly (often 45 degree counterclockwise rotated upper triangle part). While this depiction can show domain structure, it is not a reconstruction of the actual spatial structure of the chromosomes. 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 of the chromosomes. As a result we get an intuitive visual of the Hi-C contact map that not only shows the various topological domains but also how they are connected to one another. PHATE therefore provides a more ”holistic” visualization of Hi-C data compared to directly looking at the contact map.

PHATE can also effectively visualize a single chromosome. Figure 11F shows PHATE just on chromosome 1 contact map at 10 kb resolution. A 3D PHATE visualization is given in Fig. 11F, 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.

PHATE on Facebook Data reveals Super Connectors We visualize Facebook network data using PHATE. The data [47] consist of networks of friends. This network graph is directly converted into a 0-1 affinity matrix and fed to PHATE. Figure 11E 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 [47]) in Figure 13. Again, we find that PHATE finds more structure in subnetworks.

Thus, 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.

4 PHATE on Embryoid Body Data

To validate the ability of PHATE to reveal biological insights in newly measured systems, we generated single-cell RNA-sequencing data from an embryoid body differentiation system.

Embryonic stem cell (ESC) differentiation is a multi-step process that begins with the induction of primary germ layers: the ectoderm, endoderm and mesoderm. *In vitro*, primary germ layers can be induced by growing ESCs as three-dimensional aggregates, known as embryoid bodies (EB), in the absence of self-renewing signals such as TGFβ and bFGF. EB differentiation is thought to resemble embryonic development *in vivo* and has been successfully used to produce diverse, differentiated cell types, including various types of neurons, astrocytes and oligodendrocytes [48–51], hematopoietic, endothelial and muscle cells [52–60], hepatocytes and pancreatic cells [61, 62], as well as germ cells [63, 64]. However, the molecular pathways regulating germ layer development are largely unknown. It remains unclear whether *in vitro*-derived cells represent genuine functional cell types. A deeper and more systematic understanding of human ESC differentiation is necessary to overcome these challenges. Here, we begin developing such an understanding using scRNA-seq data combined with PHATE to elucidate paths of differentiation and gene-gene interactions that underlie differentiation.

We measured 31,000 cells equally distributed over a 27-day time course over which samples were collected at 3-day intervals and pooled for measurement on the 10X Chromium (see

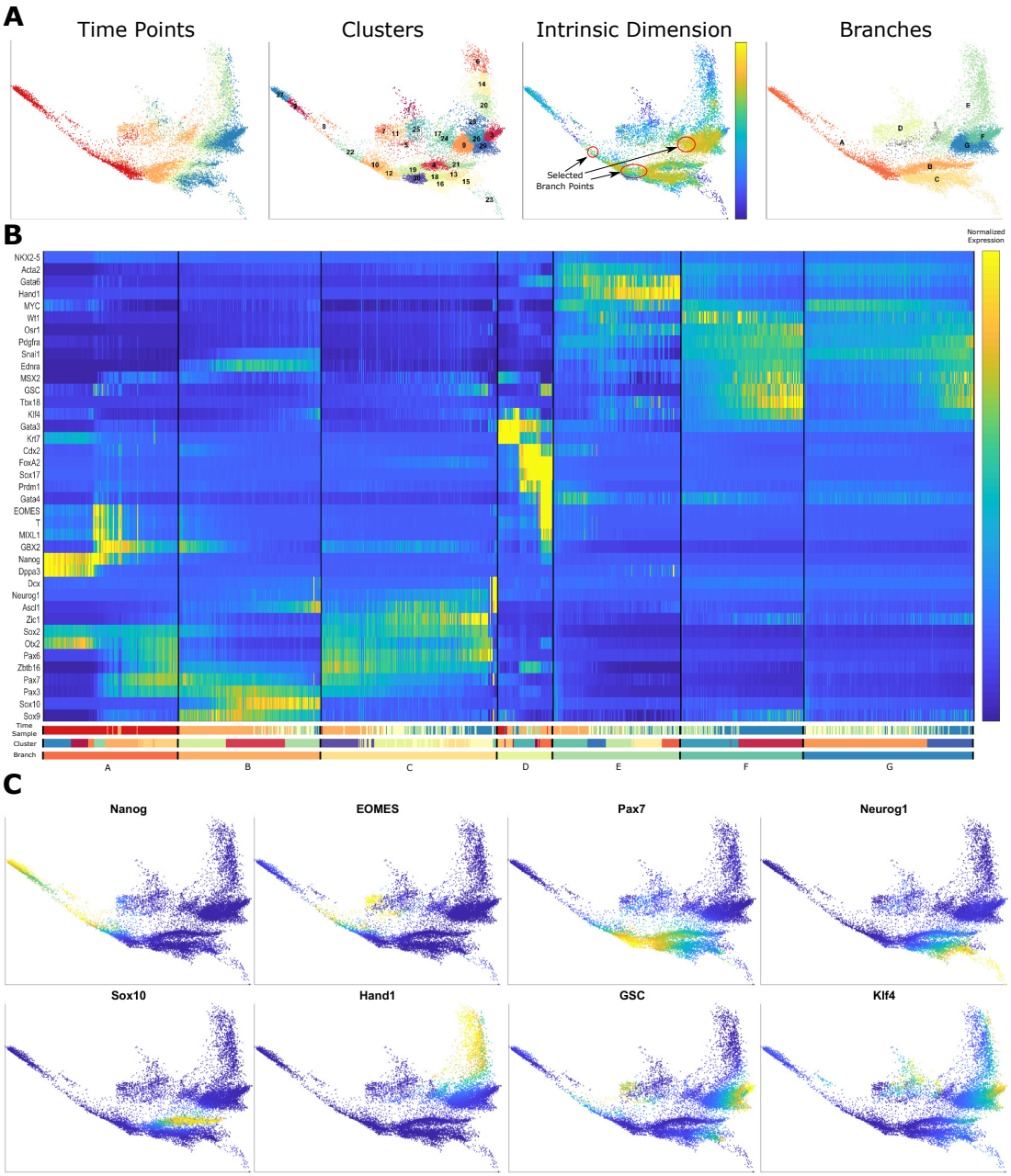

Figure 14: Analysis of hESC scRNAseq data. (**A**) Left: The PHATE visualization of the data colored by sample time. Middle left: The PHATE visualization colored by clusters. Clustering is done in the PHATE space in higher dimensions. Middle right: The PHATE visualization colored by estimated local intrinsic dimension with selected branch points highlighted. Right: Branches chosen from contiguous clusters for analysis based on the selected branch points. (**B**) Expression level of selected genes for each cell ordered by branch and ordering within the branc. Cell ordering is calculated using Wanderlust and expression levels are z-scored for each gene. Colorbars below the expression matrix identify the corresponding time sample, cluster, and branch of each cell. MAGIC is applied first to impute missing values. (**C**) The PHATE visualization colored by expression levels of selected markers.

Branch	Clusters	Cell Types
A	27, 2, 8, 22, 10, 12	Developing ESC/neuroectoderm
B	19, 4, 21	NCC
C	30, 18, 16, 13, 15, 23	NPC
D	7, 11, 25, 1	Mesoderm
E	24, 28, 20, 14, 6	Ectomesenchymal neural crest derivatives
F	26, 3	Ectomesenchymal neural crest derivatives
G	9, 29	Ectomesenchymal neural crest derivatives

Table 1: Cluster numbers from Figure 14A (left center) that comprise the branches in Figure 14A (right).

experimental methods for more details). Figure 14A (left) shows the PHATE visualization of this data, colored by time. A strong time trend is observed within the visualization, in which later time points represent greater phenotypic diversity.

4.1 Clustering Analysis

We first cluster the data to partition it into regions to verify that major cell types are present in the data. Clustering is performed using the k-means algorithm on 10 MMDS PHATE dimensions to capture the higher-dimensional structure. This can be viewed as a variation on spectral clustering using PHATE dimensions. The PHATE visualization colored by the resulting clusters is shown in Figure 14A. Clusters can be identified as known cell types based on the imputed expression of markers using the MAGIC imputation algorithm [28]. These include neurectoderm, mesoderm, neural crest and other cell types. However, this data set is dominated by progression structure and thus is not well-characterized by clusters. Therefore, instead of analyzing individual clusters, we analyze the branching progressions.

4.2 Progression Analysis

We identify progressions and branches for analysis by using the methods described in Section 2.2. The local intrinsic dimensionality is estimated on higher-dimensional PHATE dimensions to identify suggested regions as branch points (shown in Figure 14A). The centrality measure and diffusion map extrema are used to identify possible end points. We then select branches by concatenating clusters between the selected branch and endpoints based on their spacial contiguity within the visualization (shown in Figure 14A). Wanderlust [5] is then used to define an ordering of cells within each branch where the starting cell of each branch is chosen as the left-most cell.

Figure 14B shows the gene expression matrix within the identified branches for a set of well-characterized lineage-specific markers. From this matrix, we can identify different cell types and differentiation processes associated with the branches and clusters. For example,

embryonic stem cell (ESC)-specific transcripts Nanog and Dppa3 are highly expressed in cells located at the left-most part of branch A (cluster 27), indicating that this is the start 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 beginning 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 Sox9, Sox10, and Pax3, which are all associated with neural crest differentiation. In contrast, branch C shows high expression in 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 apidocyte progenitors (Snai1, Twist, Wt1, Osr1, Pdgfra, Tbx18, Acta2, Gsc, Klf4). Figure 14C shows the PHATE visualization colored by a few of these genes to highlight the different regions. This analysis demonstrates that the PHATE embedding can successfully resolve germ layers during in vitro differentiation of human ESCs.

4.3 PHATE Correlates with Known Biology

To further validate the results from our progression analysis, we compare the scRNA-seq data with bulk RNA-seq data available from the literature including datasets from embryonic stem cells (ESC) [65]; ESC-derived neuroectodermal cells (NEC) [66]; ESC differentiating into neural progenitor cells (NPC) at days 0, 8, and 16 [67]; neural crest cells (NCC) [68]; definitive endoderm cells (DEC) [65]; dental pulp stem cells (DPSC, derived from neural crest cells) [69]; human foreskin fibroblasts (HFF) [65]; and cardiomyocytes [70].

We compare the bulk and scRNA-seq data as follows. For statistical testing, we divide the cells into branches or regions of analysis as shown in Figure 15A. For each single cell, we calculate the Spearman correlation coefficient between the cell's gene expression levels (after MAGIC) and the bulk gene expression levels. In all cases, we focus only on transcription factors. Figure 15B shows the PHATE visualization colored by the resulting correlation coefficients of each cell. In this figure, we can see where the bulk RNA-seq correlates highest within the visualization. The bulk datasets ESC and NPC on day zero are highly correlated with the cells within region i (this is also visible in Figure 15A where the median correlation coefficient is calculated for each region and dataset). This corresponds to the earliest time point and is consistent with our results above. The NEC and the NPC (days 8 and 16) datasets are most highly correlated with cells in region ii, which we found from our previous analysis to be associated with the early cells in the neural crest and neural progenitors lineage. As expected from our

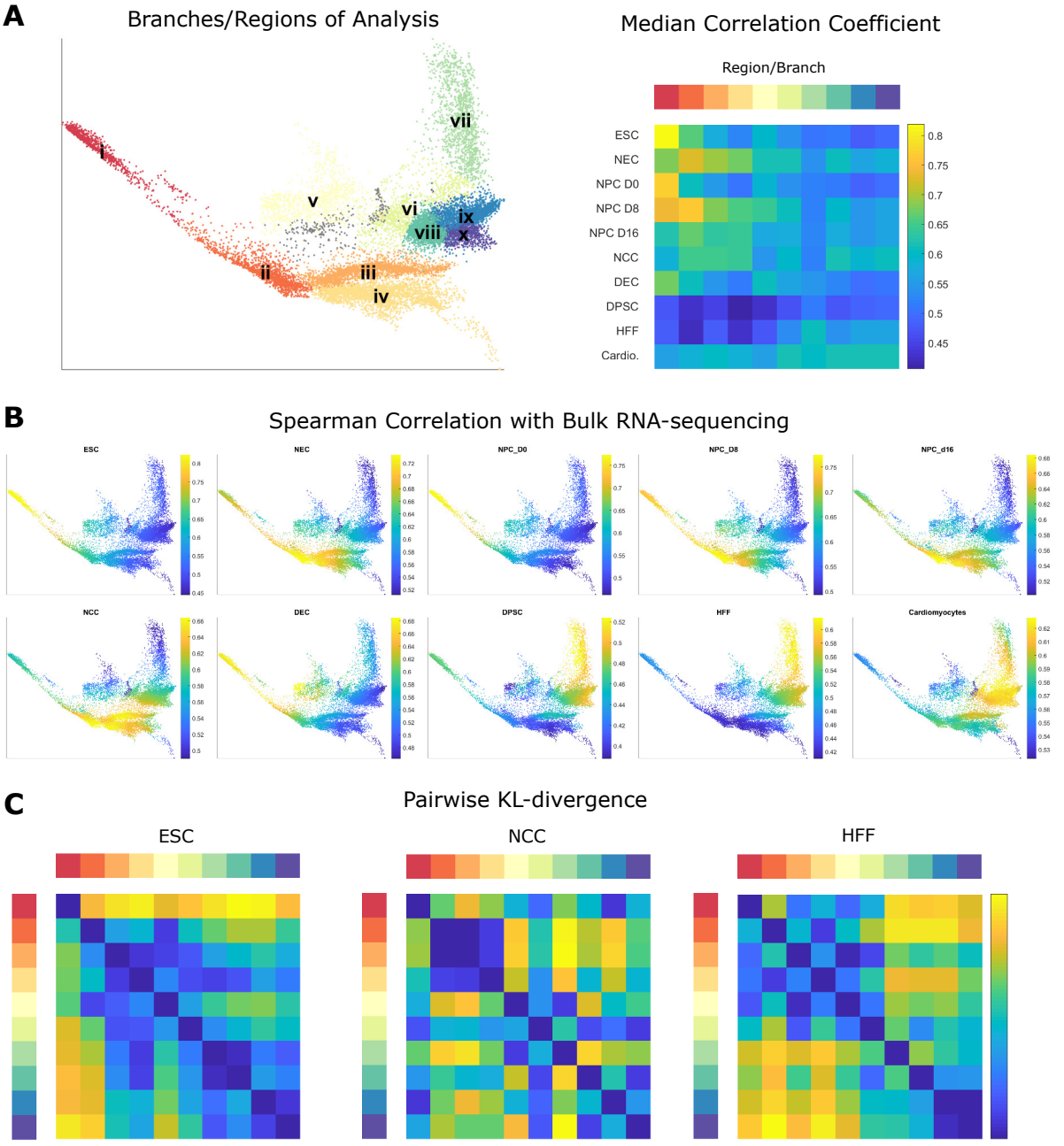

Figure 15: Comparison with bulk RNA-seq measurements. (**A**) Highlighted regions of analysis (left) and the corresponding median Spearman correlation coefficient between the cells' transcription factor expression level with the TF expression levels of the bulk RNA-seq datasets (right). (**B**) PHATE visualization with cells colored by the Spearman correlation coefficients with the bulk RNA-seq data sets. (**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 $< 3.3 \times 10^{-6}$).

previous analyses, the DEC dataset correlates the highest with cells in regions i and v while the DPSC and HFF datasets correlate the most with cells in regions vii-x, encompassing various neural crest derivatives. The cardiomyocyte dataset exhibited high correlation with the cells in regions viii, ix, and x containing cells with high expression of the smooth muscle marker Acta2. Thus the observed high correlation could reflect functional similarities between the two types of muscle progenitors other than their common origin.

We performed statistical tests on the results by computing the pairwise Kullback-Leibler (KL) divergence between the distribution of correlation coefficients for each region and for each dataset. The KL divergence was estimated using the k-nn estimator in [71]. We performed statistical tests on the pairwise divergence using the central limit theorem in [72] to test whether the estimated divergences are statistically greater than zero. 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. For illustration, some of the pairwise estimates of the KL-divergence are included in Figure 15C. The resulting matrices are consistent with the visualizations in Figure 15B. For example, for the ESC bulk data, 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. In contrast, for the NCC bulk data, the correlation coefficient is relatively high and similar in regions ii-iv. The KL divergence is relatively low in this region. Similar results were obtained when estimating the Renyi-α divergence for $\alpha = 0.5$.

Our comparison to bulk RNA-seq data provides further validation that the branches within the PHATE visualization are associated with different cell types in the EB differentiation process. This validates the utility of PHATE in extracting biological meaning from data.

4.4 PHATE Reveals Larger Developmental Signatures

Here we investigate larger developmental signatures within the EB system by examining the expression patterns within the visualization of proliferation markers, transcription factors, and chromatin modifiers.

Proliferation Markers Stem cell renewal and proliferation have been major areas of investigation recently, especially to explain tissue growth and maintenance [73]. Therefore, we examined the composite signature of multiple proliferation markers. The left plot in Figure 16A shows the average z-scored expression of 10 proliferation markers (Cdkn1a, Cdkn1b, Cdkn1c, Cdkn2a, Cdkn2b, Cdkn2c, Skp1, Ccne1, Ccne2, and MKI67). From this plot, we see that most cells express at least one of these markers, indicating that cell proliferation is occurring throughout the sample. The cells that express the proliferation markers the least are located at the endpoint of the neural crest branch (cluster 21 within branch B) and the endpoint branch F in cluster 3, suggesting that the cells in these regions are proliferating less on average than the other cells. In contrast, the cells that have the highest average proliferation marker expression levels are located in part of the neural branch (cluster 16 in branch C), the endpoint of cluster 29

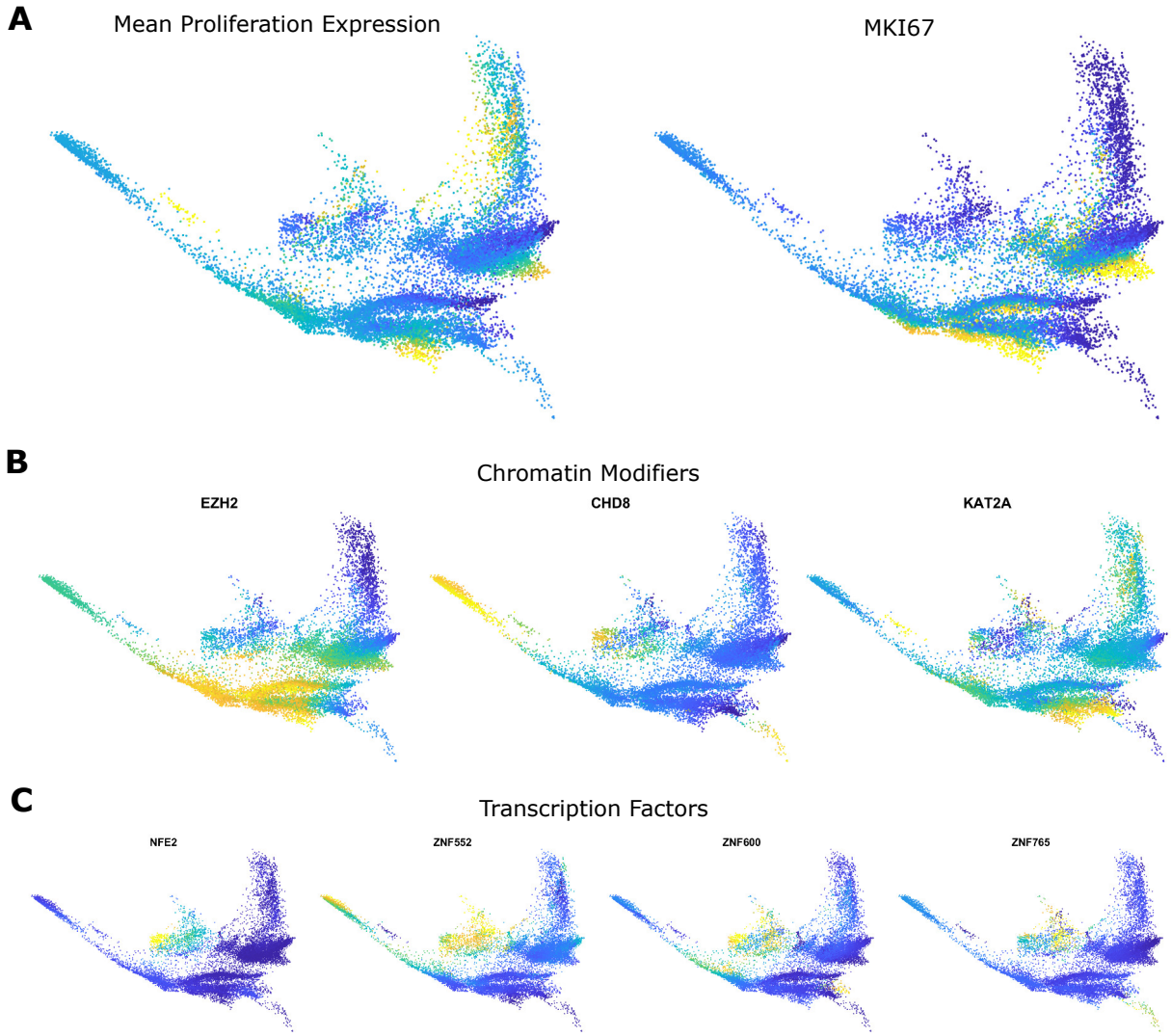

Figure 16: PHATE visualization colored by selected chromatin factors and proliferation markers. (**A**) PHATE colored by the average z-scored gene expression of 10 proliferation markers and by the general proliferation marker MKI67. (**B**) PHATE colored by selected chromatin modifiers. (**C**) PHATE colored by transcription factors that are highly expressed in cells in the mesoderm that were not associated with this layer in the literature.

in branch G, and a few other regions scattered throughout branches A, D, and E. A slightly different pattern is visible in the visualization colored by the general proliferation marker MKI67. Here the MKI67 is highly expressed in cluster 29 and in the bottom parts of the neuroectoderm and the earlier part of the neural crest (branch B) and the neural progenitor branch (branch C). With the exceptions of clusters 29, 18, and 16, MKI67 is most lowly expressed in the later time samples. However, other proliferation markers are expressed in these regions, indicating that cell proliferation is still occurring in these cell types at these times.

Chromatin Modifiers It is often thought that changes in the epigenetic state of a cell in addition to changes in the mix of transcription factors present in the cell play a key role in differentiation. Thus understanding epigenetics and specifically chromatin modification is important for understanding differentiation and also for reprogramming [74]. Figure 16B shows the PHATE visualization colored by selected chromatin modifiers. For instance, EZH2 is important for maintaining cell differentiation in ESCs [75]. EZH2 is expressed in all of the cells except for the top part in branch E and is highly expressed in the cells corresponding to neuroectoderm and the early neural and neural crest progenitors. CHD8 is important during fetal development [76] and is most highly expressed in the early stages of the data in the ESCs. KAT2A promotes transcriptional activation and is expressed in most of the cells. It is most highly expressed in parts of branch C which contains neural progenitors.

Transcription Factors Here we identify transcription factors that are expressed in cells associated with the mesoderm layer. We focus specifically on transcription factors that were not associated with the mesoderm layer in the literature. These transcription factors were identified to be among the transcription factors that are most uniquely expressed within branch D (as measured by estimating the earth-mover's distance, or EMD, between the cells in the mesoderm region and the cells in all other regions of the PHATE visualization) and that are highly expressed. Figure 16C shows four of these transcription factors. NFE2 is expressed highly in a region of branch D that also highly expresses Eomes, a gene associated with mesendoderm differentiation. Thus NFE2 may also play a role in this stage of differentiation. NFE2 is involved in regulating erythroid and megakaryocytic differentiation and maturation [77]. While these cell types are derived from the mesoderm layer, this is the first study to link NFE2 to this earlier stage of cell development.

We also identified the zinc finger proteins ZNF552, ZNF600, and ZNF765 to be highly expressed in cells in the mesoderm layer and lowly expressed in nearly all other cells. Thus these transcription factors may play a key role in the mesoderm differentiation. This demonstrates the potential of using PHATE to perform data exploration and identify new genes to identify cell populations.

5 Conclusion

The PHATE method presented in this paper provides a complete embedding and visualization of data such that its underlying structure (whether cluster-like or continuous) and progression patterns in it are revealed, while simultaneously denoising the data. This is achieved by metric embedding of a novel distance that we call potential distance. This distance is calculated by first computing a localized affinity matrix between data points, which captures local neighborhoods, then propagating affinities via a Markov diffusion operator to obtain *global* transition probabilities, and finally using a log transformation of the diffusion operator to derive an informational distance between cells. This metric has the advantage of capturing both local fine-structure and global distances within the data and enables the visualization of *transitions* in data in low-dimensional coordinates.

We propose PHATE as a key tool for data exploration in several biomedical and non-biomedical contexts. To aid data exploration, we further show that a PHATE embedding can be analyzed as a set of intersection progressions based on our method for identifying suggested branchpoints and endpoints with local intrinsic dimensionality and diffusion extrema. Then each progression in the data can be characterized in terms of its features or correlation of features along branches. We show that PHATE can be used widely in many high-dimensional datatypes such as single-cell RNA sequencing, CyTOF, image data, and connectivity data like social networks or HI-C DNA contact maps. We show that in the case of single-cell RNA sequencing data or CyTOF data, branches can be correlated by expressions of proteins or genes that may be involved in driving progressions.

We further establish the power of PHATE in exploratory data analysis by analyzing newly generated embryoid body data containing over 30,000 cells at different states of differentiation (e.g, stem cells, germ layers, and progenitor cells) collected and sequenced over a period of four weeks. Here, we demonstrated and validated that PHATE provides a comprehensive picture of lineages that are emerging from embryonic stem cells, as well as connections between such lineages. Furthermore, PHATE enables tracking of larger signatures along this comprehensive picture, including changing proliferation and chromatin modification patterns. Future work will involve testing insights gained by PHATE on embryoid body data in terms of uncharacterized transcription factors.

We expect numerous applications to benefit from the presented embedding and visualization approach of PHATE, both in high throughput genomics and, more generally, in biomedical, empirical, and computational sciences. Indeed, our results here establish PHATE as a new data analysis tool that provides one of the most crucial needs in modern data exploration - unsupervised data-driven visualization of both local and global structures in high dimensional data. It also provides an alternative to standard visualization methods such as PCA and tSNE, while enabling human-interpretable characterization of nonlinear progression patterns, which are often not well represented (if at all captured) by these methods. Further insights provided by PHATE-based data exploration in Big Data applications, together with improved scalability, will also be explored in future work.

References

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

[2] C. Trapnell, D. Cacchiarelli, J. Grimsby, P. Pokharel, S. Li, M. Morse, N. J. Lennon, K. J. Livak, T. S. Mikkelsen, and J. L. Rinn, "The dynamics and regulators of cell fate decisions are revealed by pseudotemporal ordering of single cells," *Nature biotechnology*, vol. 32, no. 4, pp. 381–386, 2014.

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] 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.

[18] 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.

[19] A. H. Rizvi, P. G. Camara, E. K. Kandror, T. J. Roberts, I. Schieren, T. Maniatis, and R. Rabadan, "Single-cell topological rna-seq analysis reveals insights into cellular differentiation and development," *Nature Biotechnology*, vol. 35, no. 6, pp. 551–560, 2017.

[20] 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.

[21] 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.

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

[23] 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.

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

[25] G. Singh, F. Mémoli, and G. E. Carlsson, "Topological methods for the analysis of high dimensional data sets and 3d object recognition," in *SPBG*, pp. 91–100, 2007.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

[41] 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.

[42] 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.

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

[44] 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.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

[49] 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.

[50] 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.

[51] 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.

[52] 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.

[53] 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.

[54] 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.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

[61] 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.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

[70] 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.

[71] 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.

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

[73] A. Mullard, "Proliferation without differentiation," *Nature Reports Stem Cells*, 2008.

[74] C. Luzzani, C. Solari, N. Losino, W. Ariel, L. Romorini, C. Bluguermann, G. Sevlever, L. Barañao, S. Miriuka, and A. Guberman, "Modulation of chromatin modifying factors' gene expression in embryonic and induced pluripotent stem cells," *Biochemical and biophysical research communications*, vol. 410, no. 4, pp. 816–822, 2011.

[75] X. Ding, X. Wang, S. Sontag, J. Qin, P. Wanek, Q. Lin, and M. Zenke, "The polycomb protein ezh2 impacts on induced pluripotent stem cell generation," *Stem cells and development*, vol. 23, no. 9, pp. 931–940, 2013.

[76] J. L. Ronan, W. Wu, and G. R. Crabtree, "From neural development to cognition: unexpected roles for chromatin," *Nature Reviews Genetics*, vol. 14, no. 5, pp. 347–359, 2013.

[77] J. J. Gasiorek and V. Blank, "Regulation and function of the nfe2 transcription factor in hematopoietic and non-hematopoietic cells," *Cellular and molecular life sciences*, vol. 72, no. 12, pp. 2323–2335, 2015.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] 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.

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

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

[85] 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.

[86] 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.

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

[88] 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.

[89] 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 provide further details on the methods and data in addition to a robustness analysis of the PHATE visualization. We first discuss the manifold and diffusion geometry data model that provides the foundation for the PHATE algorithm. We then provide a more detailed mathematical discussion of the diffusion operator including details on the α-decaying kernel and our method for choosing the time scale t. This is followed by a mathematical discussion of the potential distances and details on multidimensional scaling. We then provide details on the local intrinsic dimension estimation method from [14, 78] that we use to identify suggested branch points. The processes for generating the EB data and the artificial tree data are then described followed by details on data preprocessing methods. Finally, we demonstrate the robustness of the PHATE visualization to subsampling and the choice of the scale parameter t.

6.1 Manifold and Diffusion Geometry Data Models

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

It has been shown in several works (e.g., [79, 80]) that manifold geometries are closely related to heat diffusion, modeled by the differential heat equation, on the one hand, and to differential Laplace-Beltrami operators on the other hand. Indeed, solutions of the heat equation over a manifold capture its intrinsic properties, while providing embeddings, affinities, and distance metrics that capture intrinsic manifold relations. It has further been shown that these can

be robustly discretized for empirical observations that correlate with hidden (or latent) manifold models, e.g., by considering diffusion maps embedding of the data [10, 81, 82]. The embedding obtained by PHATE extends these results by considering an underlying geometry consisting of multiple one-dimensional manifolds (i.e., trajectory curves) that cross each other, while alleviating boundary-condition instabilities to maintain low dimensionality of the embedded space 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.

6.2 The Diffusion Operator

PHATE is based on constructing a diffusion geometry to learn and represent the shape of the data [10,81,82]. 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 $\mathcal{X} = \{x_1, \ldots, x_N\} \subset \mathbb{R}^d$ be a dataset 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 $\mathcal{M}^m \subseteq \mathbb{R}^d$ with $m \ll d$. The classic diffusion geometry proposed in [10] is based on first defining a notion of local neighborhoods in the data. A popular locality notion is given by a Gaussian kernel $k_\varepsilon(x, y) = \exp(-\|x - y\|^2/\varepsilon)$ that quantifies similarities between points based on Euclidean distances. The bandwidth ε determines the radius (or spread) of neighborhoods captured by this kernel. The kernel is then normalized with the row-sums

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

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

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

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

6.2.1 The alpha-decaying kernel and adaptive bandwidth

When applying the diffusion map framework to data, the choice of the kernel K and bandwidth ε plays a key role in the results. In particular, choosing the bandwidth corresponds to a tradeoff between encoding global and local information in the probability matrix P_ε. If the bandwidth

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

The original heuristic proposed in [10] 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 [83]. See Fig. 17 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

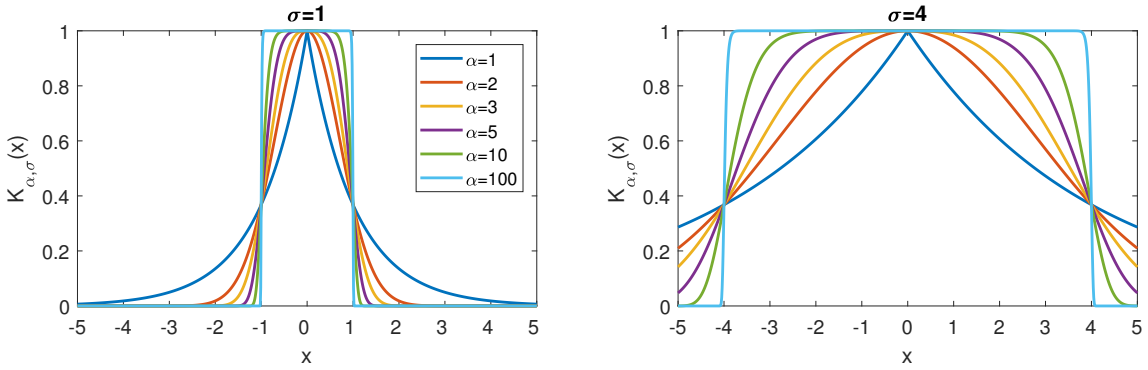

Figure 17: The α-decaying kernel $K_{\alpha,\sigma}(x) = \exp\left(-\left(\frac{|x|}{\sigma}\right)^{\alpha}\right)$ as a function of x for different values of α and $\sigma = 1$ (left) and $\sigma = 4$ (right). As α increases, $K_{\alpha,\sigma}(x)$ becomes more constant for $x \in (-\sigma, \sigma)$ and the tails of the kernel become lighter (i.e., decay to zero more quickly) for $x \notin (-\sigma, \sigma)$.

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 sparsely connected. Thus we recommend that α be fixed at a large number (e.g. $\alpha \geq 10$) and then k can be chosen to determine the connectivity of the graph. In practice, we find that choosing k to be around 5 and α to be about 10 works well for all the data sets presented in this work.

6.3 Powering the Diffusion Operator

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

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

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

For appropriate choices of kernel parameters (as 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 [82] that for fixed-bandwidth Gaussian-kernel diffusion, p^∞ converges asymptotically to the original distribution p of the data as $N \to \infty$ and $\varepsilon \to 0$.

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

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

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

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

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

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

6.3.1 Choosing the Diffusion Time Scale t with von Neumann Entropy

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

We choose the timescale t by quantifying the information in the powered diffusion operator with various values of t. This is accomplished by computing the spectral or *Von Neumann Entropy* (VNE) [84, 85] 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 beyond the "knee" of this decrease.

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

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

of the diffusion operator P with the row-sums ν. This symmetric matrix is often called the diffusion affinity matrix. 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 is given by the entropy of $\eta(t)$, i.e.,

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

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

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

To choose t, we plot $H(t)$ as a function of t as in the first column of Fig. 18. Choosing t from among the values where $H(t)$ is decreasing rapidly generally results in noisy visualizations and embeddings (second column in Fig. 18). Very large values of t result in an visualization where some of the branches or trajectories are combined together and some of the signal is lost (fourth column in Fig. 18). Good PHATE visualizations can be obtained by choosing t from among the values where the decrease in $H(t)$ is relatively slow, i.e. the set of values soon after the "knee" in the plot of $H(t)$ (third column in Fig. 18 and the PHATE visualizations in Fig. 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 the Methods section. The actual value can be chosen by selecting a t value where the second derivative of $H(t)$ is low.

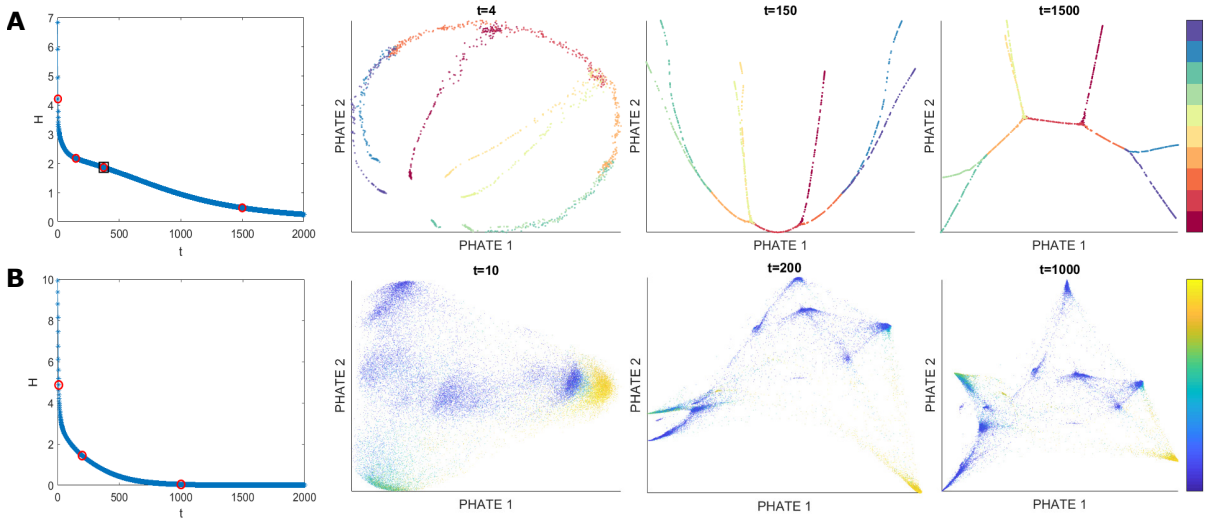

Figure 18: Demonstration of the effect of the scale t on the PHATE visualization for the (**A**) artificial tree data colored by branch and (**B**) bone marrow mass cytometry data from [86] colored by CD4 expression levels. The first column shows the VNE $H(t)$ (see Eq. 7) of the diffusion affinities 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 Figs. 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. However, the range of t values that give a good visualization is generally quite large.

6.4 Creating Potential Distances

To analyze the constructed heat diffusion process, two possible scenarios can be considered for the origin of the dataset $\mathcal{X}$ and its distribution p, as described in [81,82]. 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 [81, 82] 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 [81, 82], in both scenarios the Markov process that defines the diffusion geometry converges asymptotically to a diffusion process governed by Fokker-Planck equations with a potential $2U(x)$, whether the original potential is defined naturally or artificially.

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

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

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

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

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

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

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

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

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

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

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

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

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

which yields the result in the proposition. □

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

6.5 Diffusion Potential Embedding with MDS

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

Some common distance embedding methods are known as multidimensional scaling (MDS). Classical MDS [87] takes a distance matrix as input and embeds the data into a lower-dimensional space using eigendecomposition techniques. We apply classical MDS to the potential distances of the data to obtain an initial configuration of the data in low dimension.

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:

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

over embedded d'-dimensional coordinates $\hat{x}_i \in \mathbb{R}^{d'}$ 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 d' is too small to represent the data without distortion. Thus, by choosing the number of MDS dimensions to be $d' = 2$ (or $d' = 3$) for visualization purposes, we trade off distortion in exchange for readily visualizable coordinates. However, 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 applied to higher values of d'.

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. However, 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.

6.6 Local Intrinsic Dimension Estimation Details

Here we provide details on the local intrinsic dimension estimation method derived in [14, 78] that we use. 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

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

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

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

where $\beta(m) = (m - \gamma)/m$, ϵ_n is an error term that decreases to 0 as $n \to \infty$, and c is a constant with respect to $\beta(m)$ [78]. 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 [14, 78].

A local estimator of intrinsic dimension $\tilde{\mathbf{m}}(i)$ at a point $\mathbf{z}_i$ can be defined by running the above procedure in a smaller neighborhood about $\mathbf{z}_i$. This approach is demonstrated in Fig. 5A, where a k-nn graph is grown locally at each point in the data. However, this estimator can have high variance within a neighborhood. To reduce this variance, majority voting within a neighborhood of $\mathbf{z}_i$ can be performed:

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

where $\mathbb{1}(\cdot)$ is the indicator function [14].

6.7 Generation of Human EB Data

Low passage H1 hESCs were maintained on Matrigel-coated dishes in DMEM/F12-N2B27 media supplemented with FGF2. For EB formation, cells were treated with Dispase, dissociated into small clumps and plated in non-adherent plates in media supplemented with 20% FBS, which was prescreened for EB differentiation. Samples were collected during 3-day intervals during a 27 day-long differentiation timecourse. An undifferentiated hESC sample was also included (Fig. 19). 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.

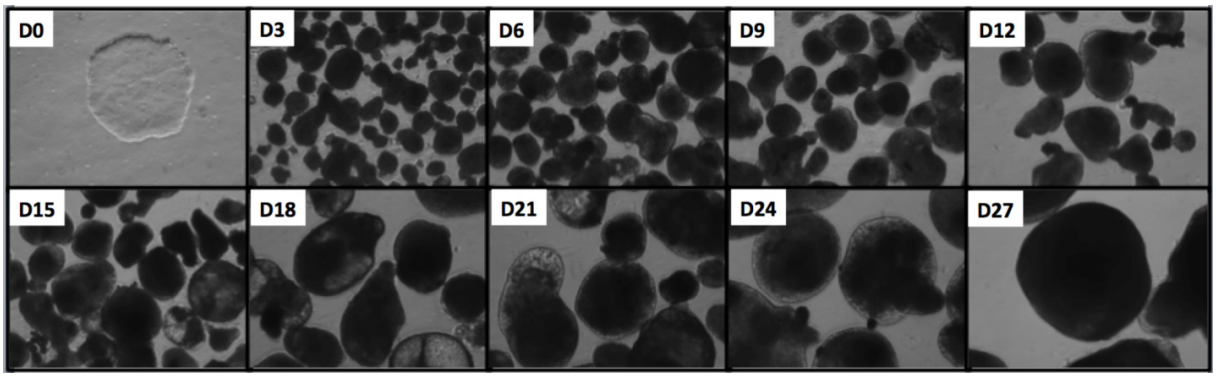

Figure 19: Inverted images of hESCs and EBs at each timepoint of data collection. Structures of different densities are clearly visible late in the time course (D15-D27) indicating the formation of distinct cell types.

6.8 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. 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.

6.9 Data Processing

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

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

Mass Cytometry Data Preprocessing We process the mass cytometry datasets according to [86].

Single-Cell RNA-Sequencing Data Preprocessing This data was processed from raw reads to molecule counts using the Cell Ranger pipeline [88] 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 [89]. 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 scRNAseq 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 in Figure 1D.

6.10 Robustness Analysis of PHATE

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

Robustness to Subsampling We demonstrate that the PHATE algorithm is robust to subsampling of the data by running PHATE on the mass cytometry bone marrow dataset from [86] with varying subsample sizes N. The PHATE embedding for $N = 30000$ is shown in Figure 18B while Figure 20 shows the PHATE embedding for $N = 1000, 2500, 5000, 10000$. Note that

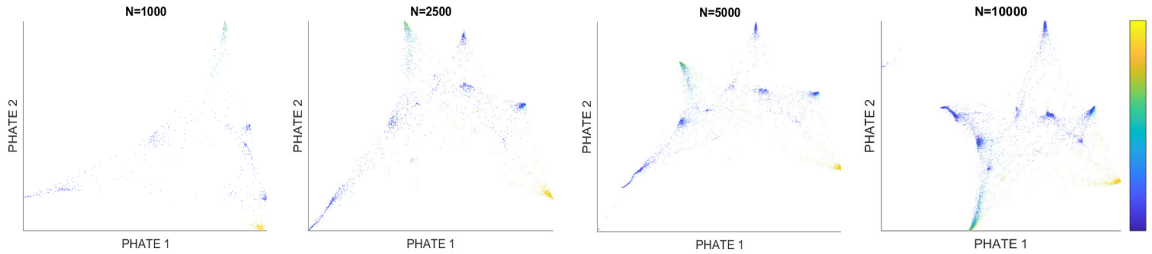

Figure 20: The PHATE visualization for the bone marrow mass cytometry dataset from [86] with varying number of subsample sizes N. The coloring corresponds to CD4 expression level. Most branches present for $N = 10000$ are also visible when $N = 5000$ or $N = 2500$ while several branches are still visible for even $N = 1000$, demonstrating that the PHATE embedding is robust to the size of the subsample. See also Figure 18B for $N = 30000$.

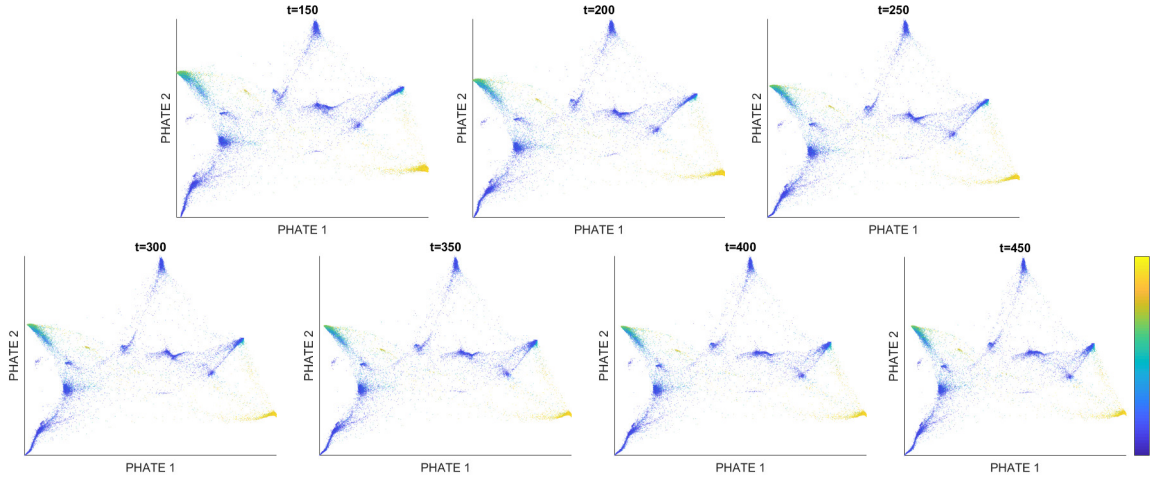

Figure 21: The PHATE visualization for the bone marrow mass cytometry dataset from [86] with varying scale parameter t. The embeddings for all t preserve the branching structure and the visualizations are nearly identical to each other. This demonstrates that the embedding is robust to the choice of t.

most of the branches or trajectories that are visible when $N = 30000$ are still visible when $N = 10000, 5000$, and 2500. Even when $N = 1000$, several branches are still visible in the embedding. Thus, PHATE is robust to the subsampling size. Similar results can be obtained on other datasets.

Robustness to t Here, we show that the PHATE embedding is quite robust to the choice of t. Figure 21 shows the PHATE embedding on the bone marrow mass cytometry dataset from [86] with varying scale parameter t. Figure 21 shows that the embeddings for $85 \leq t \leq 115$ are nearly identical. Additionally, the embeddings for $t = 50$ and $t = 150$ are very similar to the embedding for $t = 100$. Thus, PHATE is also very robust to the scale parameter t. Similar results can be obtained on other datasets.

Software Software for PHATE is available via github for academic use: https://github.com/KrishnaswamyLab/PHATE.