

# Multiscale PHATE Exploration of SARS-CoV-2 Data Reveals Multimodal Signatures of Disease

**Manik Kuchroo**

Yale School of Medicine

**Jessie Huang**

Yale School of Medicine

**Patrick Wong**

Yale School of Medicine

**Jean-Christophe Grenier**

Montreal Heart Institute

**Dennis Shung**

Yale School of Medicine

**Alexander Tong**

Yale University

**Carolina Lucas**

Yale University <https://orcid.org/0000-0003-4590-2756>

**Jon Klein**

Yale University <https://orcid.org/0000-0002-3552-7684>

**Daniel Burkhardt**

Yale University <https://orcid.org/0000-0001-7744-1363>

**Scott Gigante**

Yale University <https://orcid.org/0000-0002-4544-2764>

**Abhinav Godavarthi**

Yale College

**Ben Israelow**

Yale University

**Tianyang Mao**

Yale School of Medicine

**Ji Eun Oh**

Yale School of Medicine

**Julio Silva**

Yale School of Medicine

**Takehiro Takahashi**

Yale School of Medicine

**Camila Odio**

Yale School of Medicine  
**Arnau Casanovas-Massana**  
Yale University

**John Fournier**  
Yale School of Medicine

**Shelli Farhadian**  
Yale University <https://orcid.org/0000-0001-7230-1409>

**Charles Dela Cruz**  
Yale University

**Albert Ko**  
Yale University

**Matthew Hirn**  
Michigan State University <https://orcid.org/0000-0003-0290-4292>

**Francis Wilson**  
Yale School of Medicine

**Julie Hussin**  
Montréal Heart Institute

**Guy Wolf**  
University of Montreal <https://orcid.org/0000-0002-6740-059X>

**Akiko Iwasaki**  
Yale University <https://orcid.org/0000-0002-7824-9856>

**Smita Krishnaswamy** (✉ [smita.krishnaswamy@yale.edu](mailto:smita.krishnaswamy@yale.edu))  
Yale University <https://orcid.org/0000-0001-5823-1985>


---

## Article

**Keywords:** SARS-CoV-2, COVID-19, Multiscale PHATE

**Posted Date:** March 15th, 2021

**DOI:** <https://doi.org/10.21203/rs.3.rs-311045/v1>

**License:**  This work is licensed under a Creative Commons Attribution 4.0 International License.  
[Read Full License](#)

---

**Version of Record:** A version of this preprint was published at Nature Biotechnology on February 28th, 2022. See the published version at <https://doi.org/10.1038/s41587-021-01186-x>.

# Multiscale PHATE Exploration of SARS-CoV-2 Data Reveals Multimodal Signatures of Disease

Manik Kuchroo<sup>1\*</sup>, Jessie Huang<sup>2\*</sup>, Patrick Wong<sup>3\*</sup>,  
Jean-Christophe Grenier<sup>4</sup>, Dennis Shung<sup>5</sup>, Alexander Tong<sup>2</sup>, Carolina Lucas<sup>3</sup>, Jon Klein<sup>3</sup>,  
Daniel B. Burkhardt<sup>6</sup>, Scott Gigante<sup>7</sup>, Abhinav Godavarthi<sup>8</sup>, Bastian Rieck<sup>9</sup>, Benjamin  
Israelow<sup>3</sup>, Michael Simonov<sup>5</sup>, Tianyang Mao<sup>3</sup>, Ji Eun Oh<sup>3</sup>, Julio Silva<sup>3</sup>, Takehiro Takahashi<sup>3</sup>,  
Camila D. Odio<sup>5</sup>, Arnau Casanovas-Massana<sup>10</sup>, John Fournier<sup>11</sup>, Yale IMPACT Team<sup>12</sup>,  
Shelli Farhadian<sup>11</sup>, Charles S. Dela Cruz<sup>13</sup>, Albert I. Ko<sup>10</sup>, Matthew J. Hirn<sup>14,15</sup>, F. Perry  
Wilson<sup>16</sup>, Julie Hussin<sup>4,17§</sup>, Guy Wolf<sup>18,19§</sup>, Akiko Iwasaki<sup>3,20§,†</sup> and Smita Krishnaswamy<sup>2,6§,†</sup>

<sup>1</sup>Department of Neuroscience, Yale University, New Haven, CT, <sup>2</sup>Department of Computer Science, Yale University, New Haven, CT, <sup>3</sup>Department of Immunobiology, Yale University, New Haven, CT, <sup>4</sup>Montreal Heart Institute, Montréal, Québec, Canada, <sup>5</sup>Department of Medicine, Yale University, New Haven, CT, <sup>6</sup>Department of Genetics, Yale University, New Haven, CT, <sup>7</sup>Computational Biology, Bioinformatics Program, Yale University, New Haven, CT, <sup>8</sup>Department of Applied Mathematics, Yale University, New Haven, CT, <sup>9</sup>Department of Biosystems Science and Engineering, ETH Zurich, Switzerland, <sup>10</sup>Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT, <sup>11</sup>Department of Medicine, Section of Infectious Diseases, Yale University School of Medicine, New Haven, CT, <sup>12</sup>A list of authors and their affiliations appears at the end of the paper, <sup>13</sup>Department of Medicine, Section of Pulmonary and Critical Care Medicine, Yale University School of Medicine, New Haven, CT, <sup>14</sup>Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI, <sup>15</sup>Department of Mathematics, Michigan State University, East Lansing, MI, <sup>16</sup>Clinical and Translational Research Accelerator, Department of Medicine, Yale University, New Haven, CT, <sup>17</sup>Faculty of Medicine, Université de Montréal, Québec, Canada, <sup>18</sup>Mila – Quebec AI institute, Montréal, Quebec, Canada, <sup>19</sup>Department of Mathematics and Statistics, Université de Montréal, Montréal, Quebec, Canada, <sup>20</sup>Howard Hughes Medical Institute, Chevy Chase, MD, USA.

\* Equal contribution. § Jointly supervised work.

† Correspondence to Smita Krishnaswamy, 333 Cedar St, New Haven, CT 06520. E-mail: [smita.krishnaswamy@yale.edu](mailto:smita.krishnaswamy@yale.edu); and to Akiko Iwasaki, 300 Cedar St, New Haven, CT 06520. E-mail: [akiko.iwasaki@yale.edu](mailto:akiko.iwasaki@yale.edu).

# 1 Abstract

The biomedical community is producing increasingly high dimensional datasets, integrated from hundreds of patient samples, which current computational techniques struggle to explore. To uncover biological meaning from these complex datasets, we present an approach called Multiscale PHATE, which learns abstracted biological features from data that can be directly predictive of disease. Built on a coarse graining process called diffusion condensation, Multiscale PHATE learns a data topology that can be analyzed at coarse levels for high level summarizations of data, as well as at fine levels for detailed representations on subsets. We apply Multiscale PHATE to study the immune response to COVID-19 in 54 million cells from 168 hospitalized patients. Through our analysis of patient samples, we identify  $CD16^{hi}CD66b^{lo}$  neutrophil and  $IFN\gamma^{+}GranzymeB^{+}$  Th17 cell responses enriched in patients who die. Furthermore, we show that population groupings Multiscale PHATE discovers can be directly fed into a classifier to predict disease outcome. We also use Multiscale PHATE-derived features to construct two different manifolds of patients, one from abstracted flow cytometry features and another directly on patient clinical features, both associating immune subsets and clinical markers with outcome.

# 2 Introduction

Extremely high throughput biomedical data is generated by a range of technologies [1–6] that measure dozens to tens of thousands of features in millions of individual cells. Furthermore, these technologies are now applied to large patient cohorts, providing information that must be integrated and analyzed at scale to provide insights into cellular mechanisms and patient responses. However, there are no specific methods designed to sift through such data at varying levels of granularity to uncover features that are directly associated with disease phenotype. The SARS-CoV-2 pandemic has brought this problem to the forefront of biologists’ minds. As increasingly large datasets are built by integrating patient samples from around the globe, computational approaches also must scale to provide improved insights regardless of technology type.

We posit here that the key to understanding such vast and complex data is to create meaningful representations that uncover structure at all resolutions or scales. This approach involves learning representations of the biological system at many levels, allowing for coarse, high level summarization as well as fine grained, detailed representations of data subsets. Current tools for dimensionality reduction and data exploration - including t-distributed stochastic neighborhood embedding (tSNE) [7], uniform manifold approximation and projection (UMAP) [8], as well as principle component analysis (PCA) [9] - only show a single level of granularity of the data. Recent computational papers on SARS-CoV-2 have represented data using one of these approaches [10, 11], visualizing the major cell types such as B cells, T cells and myeloid cells. Differences between an effective immunological response and an ineffective one, however, may not be found at the granularity of immune compartment abundance alone. In fact, appreciation of a finer resolution of the T cell manifold would reveal subsets that may be predictive of disease severity. This phenomenon is found across biomedical data science, as the state space of the data is generally a collection of manifolds or continuum structures which can be organized at varying levels of hierarchy.

Based on this insight, we developed Multiscale PHATE, a method that can learn and visualize abstract cellular features and groupings of the data at *all levels of granularity*. Our algorithm is based on a dynamic process we have developed called diffusion condensation [12], which computes a manifold-intrinsic diffusion space on the original data before slowly condensing data points towards local centers of gravity to form natural, data-driven groupings across multiple granularities. This

coarse graining process learns the topology of the underlying dataset not by forcing merges at each iteration, as done in most agglomerative hierarchical clustering methods, but by allowing cells to naturally come together over the course of successive condensation steps. Visualizing a series of iterations in this dynamic condensation process using the manifold-affinity preserving PHATE method creates Multiscale PHATE embeddings, while evaluating connected cells across granularities creates Multiscale PHATE clusters. Furthermore through efficient scalable implementation, we show that we are able to perform condensation, visualization and clustering of large-scale the data significantly faster than "single-scale" visualization techniques like tSNE, UMAP or PHATE [13].

We showcase our method on 251 blood samples from 168 patients infected with SARS-CoV-2 measured across four different flow cytometry panels, an expanded iteration of a previously published dataset [14]. Analysis of 54 million of cells by single resolution dimensionality reduction and clustering algorithms would take days to weeks to perform. With our unique multigranular approach, we can produce high level summarizations as well as detailed cell type specific analyses of each panel of markers within minutes. When combined with the MELD [15], we find that our approach is particularly powerful at identifying canonical and non-canonical cellular populations associated with patient outcome across resolutions. At coarse resolution, we identify T cells to be broadly protective while monocytes and granulocytes to be pathogenic. At finer resolution, we identify unique non-canonical populations of cells, such as  $D16^{hi}CD66b^{-}$  neutrophil,  $CD14^{-}CD16^{hi}HLA-DR^{lo}$  monocytes, and  $IFN\gamma^{+}GranzymeB^{+}$  Th17 cells, to be associated with patient mortality. This type of multigranular analysis reveals that though broadly a cell type, such as T cells, may be protective, fine grain analysis reveals cellular subsets that can be pathogenic, highlighting the need for a multiresolution approach. Next, we show that Multiscale PHATE-derived cellular groupings can be used as features input to a random forest classifier to predict outcome better than immunologist curated and gated features. A unique contribution we make is the use of these multiscale feature proportions as descriptors of each patient, which can be used to create a patient-level embedding.

Finally, to display the generalizability of our approach across data types, we created a multiscale distillation of clinical data from 2,135 patients admitted to Yale New Haven Hospital (YNHH). Built from 18 laboratory, clinical, and demographic variables, Multiscale PHATE was able to create multiresolution embeddings of patient clinical states and identify regions enriched for different patient outcomes. By associating clinical features and cellular populations with outcomes, we found markers of multi-organ dysfunction to be associated with mortality and overall T cell counts to be associated with length of recovery from infection.

## 3 Results

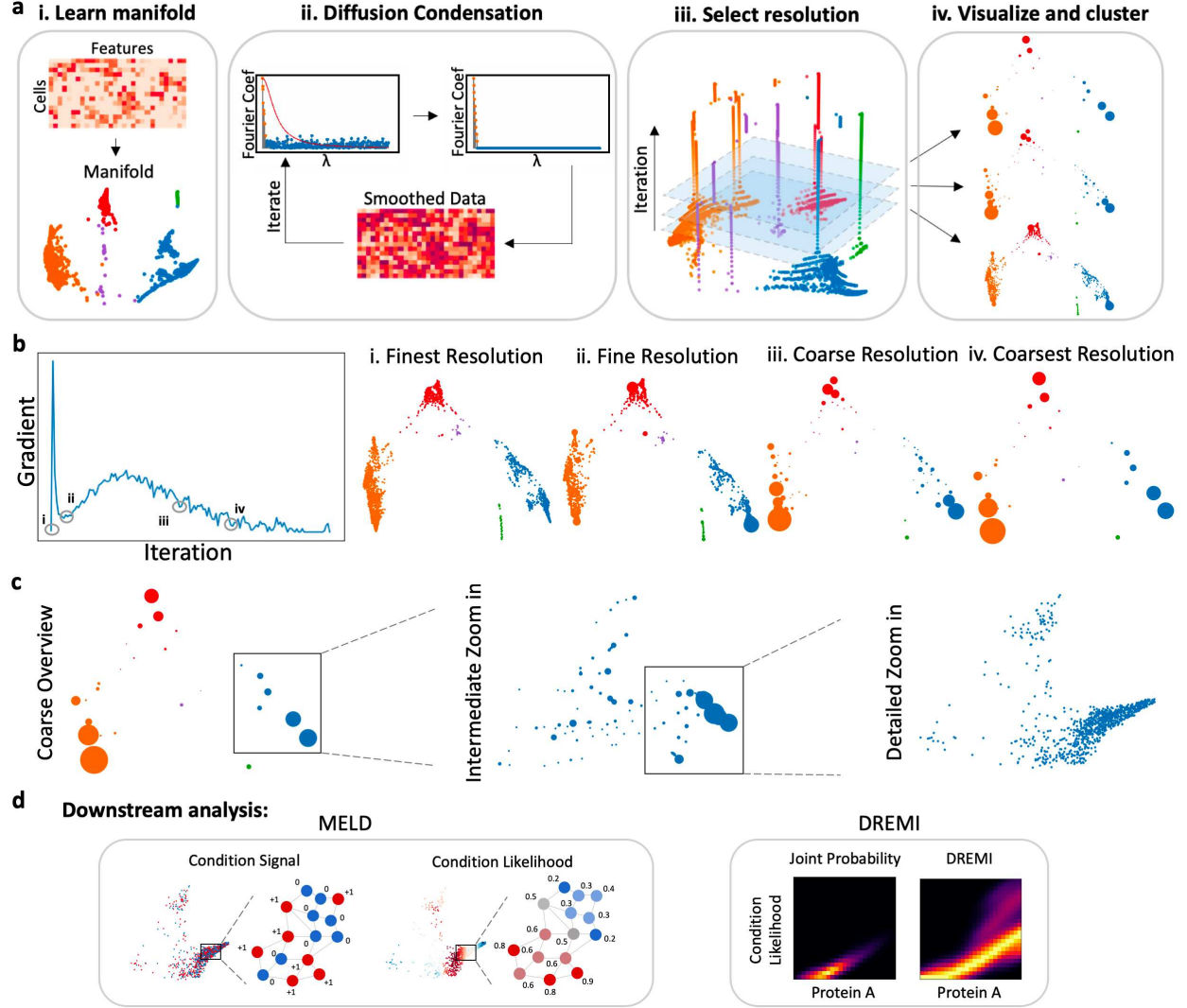
### 3.1 Overview

Although biomedical data is being generated in increasingly larger volumes, current dimensionality reduction approaches only offer a single-level view of the data. This can be problematic for several reasons. First, biomedical systems are naturally hierarchical and structure can exist at many levels of resolution—thus a single level of resolution can miss predictive features of the data. Second, as the sheer volume of data increases, most dimensionality reduction methods suffer from crowding problems where it can be difficult to ascertain density and overall structure of the data. In order to address these issues we develop multiscale PHATE which offers higher level summarizations of the data along with the ability to "zoom in" to regions to reveal additional structure and detail.

Multiscale PHATE utilizes a data coarse graining approach known as diffusion condensation that learns the manifold geometry of the data across all levels of granularity [12]. We apply diffusion condensation starting with the original data in a manifold coordinate space. The condensation

process iteratively moves points to the average of their diffusion neighbors eventually creating clusters of many levels of resolution. Then Multiscale PHATE utilizes a dimensionality reduction method optimized for structure-preserving visualization called PHATE [13] to visualize select levels.

### 3.2 Multiscale PHATE Algorithm



**Figure 1: Overview of Multiscale PHATE algorithm**

a. Multiscale PHATE process involves four successive steps. The first step (i) learns the manifold geometry via diffusion potential calculation. The second step (ii) iteratively coarse grains the manifold construction through a fast diffusion condensation process to learn data topology. The third step (iii) involves the selection of salient granularities via gradient analysis before finally visualizing and clustering the manifold in the fourth step (iv).

b. Gradient analysis identifies a range of scales for visualization.

c. Multiscale PHATE allows for high level summarizations of data as well as finer grain zoom ins of data subsets for additional detail.

d. Multiscale PHATE abstractions of data are amenable to downstream analyses with algorithms like MELD [15] and DREMI [16].

Multiscale PHATE is a multiresolution dimensionality reduction and clustering method that can reveal the structure of data visually and quantitatively at multiple granularities. Multiscale PHATE is unique in that it is a topological data analysis method that is based in data diffusion geometry.

While geometric analysis on a cellular manifold is based on *distances* between data points, topological analysis is based on how these points relate to one another in their local geometries. Multiscale PHATE, as a tool based on concepts from *topological data analysis*, constitutes a hybrid framework between a purely geometrical view and a purely topological one [17]. In classical topological data analysis, topology is learned by first computing a pairwise distance matrix  $\mathbf{D}$ , and then identifying all point pairs whose distance falls below a distance threshold  $\delta$ . A pair of cells that fall below this threshold are deemed ‘connected.’ The value of  $\delta$  is akin to a granularity level at which we view the data: as  $\delta$  increases, more point pairs will become connected, quickly creating more connected components at coarser granularities. Finally, this process stops when  $\delta$  is sufficiently large such that all points are connected. This process of iterated merges reveals the structure of the underlying dataset across granularities by first identifying how cells relate to one another in their local geometries at small values of  $\delta$  before seeing how more dissimilar cells relate to one another at large values of  $\delta$ .

Applying these concepts to biological datasets, however, has proved to be problematic. Biomedical data has been shown to be highly non-linear, requiring non-linear diffusion-based approaches to visualize and analyze biological manifolds [13, 18]. Classical topological data analysis operates in the ambient measurement space and is incompatible with more complex manifolds. For this reason, we use diffusion condensation, which learns the topology of complex manifolds [12] in the manifold geometric coordinates [13].

Diffusion condensation learns data topology across granularities through a dynamic coarse graining process by which data points slowly and iteratively come together at a rate determined by the diffusion probabilities between them [12]. This iterative process is powerful and intuitively relates to topological data analysis, as it reveals structure and groupings of the data at all levels of granularity. The diffusion condensation process involves three steps that are repeated until all points converge:

1. Compute a Markov diffusion operator from the data
2. Apply operator to the data as a low-pass filter, moving points towards local centers of gravity
3. Merge points together that fall below a preset distance threshold  $\zeta$  to create connected components, as done in topological data analysis

**Step 1** is performed by computing a distance matrix  $\mathbf{D}$  between all data points, similar to topological data analysis. We then use a fixed bandwidth Gaussian kernel function to convert the distance matrix  $\mathbf{D}$  into an affinity matrix  $\mathbf{K}$ , so that similarity between two cells decreases exponentially with their distance as done previously [12].  $\mathbf{K}$  is then row normalized to obtain a *diffusion operator* between data points, representing the probability distribution of transitioning from one point to another in a single step.

**Step 2** The diffusion operator is applied (left-multiplied) to the input data, effectively replacing the value of a point with the weighted average of its diffusion neighbors. This process simulates a one step transition of every cell to its diffusion neighbors, causing points to *condense*, or move closer together due to the removal of variation. In graph spectral terms, this step is akin to applying a *low-pass filter* on the graph frequency spectrum which smooths the data.

**Step 3** the third step, diffusion condensation merges points that have condensed within a preset merge threshold  $\zeta$  together.



These steps are then repeated iteratively. At each step, diffusion operator is calculated on the output of the previous iteration and used as a low-pass filter to produce an increasingly coarse dataset for the next iteration. Over the course of many iterations, data points slowly converge to local centers of gravity and collapse into each other, effectively creating connected components at that level of granularity. As the diffusion condensation process continues, we scale the diffusion operator to diffuse points to more global centers of gravity. This first removes local variability in the data at initial iterations before removing more global variability in later iterations. This deep cascade of low pass filters effectively builds a topological understanding of the data by merging data points together in a natural manner across granularities.

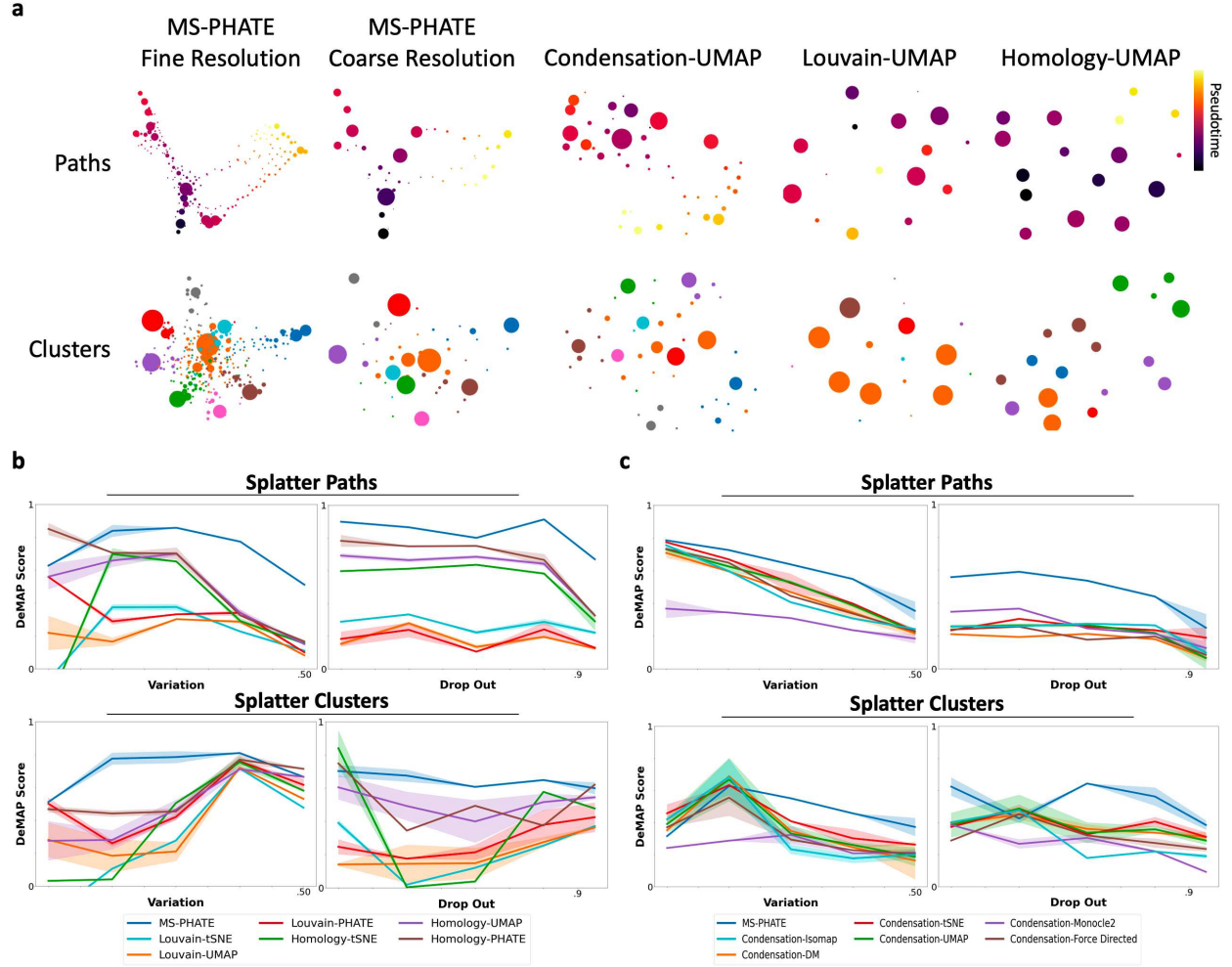
In its original form, however, the diffusion condensation process does not scale to millions of data points, does not condense points on a manifold, and is not optimized for visualization. Thus, we have modified diffusion condensation to allow for scalable and effective visualizations. The main steps of Multiscale PHATE include:

1. Computing a *fast* diffusion condensation process that scales to millions of data points,
2. Transforming datapoints to a novel diffusion potential coordinate system to learn the data manifold,
3. Identifying levels for visualization based on gradient analysis and creating a density aware visualization.

Calculation of the complete topology via diffusion condensation is computationally expensive on massive single cell datasets currently being produced. In order to allow multiscale PHATE to scale to millions of cell, we apply an initial coarse graining step before running diffusion condensation. This initial coarse graining step reduces the numbers of points while maintaining data geometry potentially by two orders of magnitude. Running diffusion condensation in ambient measurement space results in averaged points off of highly curved manifolds, such as biological ones (Extended Data Figure 1a). As a result, diffusion condensation as described above is not amenable to learning or visualizing high dimensional cellular data across granularities. In order to learn the underlying manifold, we compute manifold dimensions in the form of diffusion potential as done in PHATE [13]. This not only allows for condensation of points on the manifold but also easy visualization of intermediate granularities. Finally, in order to identify meaningful resolutions for downstream analysis, we perform retrospective gradient analysis on the diffusion condensation process to identify metastable states (Figure 1b), allowing for high level summarizations as well as zoom ins on important data subsets. Finally, to create a density aware visualization, points are sized by the number of original cells that have merged to create that connected component (Figure 1c).

Each of these steps are explained in further detail in methods (see Methods).





**Figure 2: Comparison of Multiscale PHATE with other dimensionality reduction tools**

*a. Visual comparison of Multiscale PHATE with other multiscale dimensionality reduction tools on synthetic single cell data with either path or cluster structure.*

*b. Quantitative study comparing embeddings produced by Multiscale PHATE and visualization strategies which either employ community based or topologically based abstractions of data. Comparisons were evaluated using DeMAP with increasing levels of 2 different types of biological noise, drop out and variation, as well as on data with different structures, clusters and paths. Shading represents standard deviation around mean DeMAP score for each comparison.*

*c. Quantitative study comparing embeddings produced by Multiscale PHATE and visualization strategies which visualize condensation based abstractions of data. Comparisons were evaluated using DeMAP with increasing levels of 2 different types of biological noise, drop out and variation, as well as on data with different structures, clusters and paths. Shading represents standard deviation around mean DeMAP score for each comparison.*

**Multiscale PHATE embeddings preserve local and global distances better than other multiscale visualization approaches.**

In order to quantify the quality of our dimensionality reduction strategy compared to other multiscale implementations of established visualization tools, we computed Denoised Manifold Affinity Preservation (DeMAP) scores [13] on embeddings of a variety of splatter simulated datasets [19]. While there is information loss in any sort of dimensionality reduction technique, an ideal embedding should capture as much local and global distance information as possible. To judge both local and

global distance preservation, DeMAP quantifies the ability of an embedding to preserve ground truth manifold distances, also known as geodesic distances, in a low dimensional visual representation.

In our comparisons, we performed two different ablation studies to determine the necessity of both the diffusion condensation approach to learning data topology (Figure 2b) as well as PHATE to learn manifold geometry (Figure 2c). In each study, we repeated comparisons on a variety of datasets which have different geometries, either paths (or trajectories) or cluster structure with increasing amounts of two types of biological noise: variation and dropout. After visualizing synthetic single cell datasets produced by splatter (Figure 2a) and running all comparisons, Multiscale PHATE performed superior to other methods across nearly all ranges of biological noise (Figure 2b,c). In particular, upon visual and quantitative comparison, multiscale PHATE performed superior to all other methods when visualizing trajectory structures in low dimensional embeddings (Figure 2a,b,c).

### **Multiscale clusters more accurately capture established groupings of data in synthetic and real biological datasets.**

In order to quantify the clustering accuracy of Multiscale PHATE on increasingly noisy and multigranular data, we simulated two and three-layer hierarchical stochastic block models (SBM) (Extended Data Figure 3a). In these models, a graph is constructed in which there are coarse grain clusters, each of which could be further broken down into increasingly granular clusters. In order to compare all clustering techniques across a range of noise levels, increasing amounts of random gaussian noise is added to the edge weights of the graph. At each level of noise, cluster labels are computed with multiple clustering tools: Multiscale PHATE, Louvain [20], Leiden [21] and single linkage hierarchical clustering [22]. These comparisons are run on a range of noise levels and replicated 10 times across a range of initial hierarchical SBM edge weights in both the two-layer and the three-layer models. Across both models, Multiscale PHATE performed superior to other hierarchical clustering techniques in 35 of the 42 comparison conditions (Extended Data Figure 3b,c), with only poorer performance at the finest granularity of the 3 layer SBM. Finally, in order to determine the clustering accuracy of Multiscale PHATE across granularities on real biological data, we applied our approach to flow cytometry data where cell type labels have already been established. Taking the cell population labels as identified with conventional gating analysis, we computed ARI scores on clustering outputs from a range of clustering techniques at multiple granularities. Across both fine and coarse grain clusters, Multiscale PHATE computed clusters that more faithfully represented the underlying known biological cell types (Extended Data Figure 3d).

### **Generalizability, scalability and reproducibility of Multiscale PHATE**

Multiscale PHATE is broadly generalizable to a vast number of biological data types including flow cytometry, scRNAseq, scATACseq, clinical variables among others (Extended Data Figure 2). When comparing run times between different techniques, it becomes clear that Multiscale PHATE is able to rapidly scale to millions of cells, successfully embedding 5 million cells in less than 10 minutes, while the next most scalable technique, Monocle2, can only embed 500,000 cells in a comparable time (Extended Data Figure 1c). Across all comparisons, it is important to note that number of features did not alter run time drastically. Biological technologies applied to single cells measure tens to hundreds of thousands of features. Since the initial step of each of these dimensionality reduction algorithms is feature compression with PCA, the only major difference in run time will be length to compute PCA compression of data. With Randomized SVD, this process is rapid for a number of features lengths, leading to similar run times across technology types. Finally, Multiscale PHATE is highly reproducible. A common issue with UMAP and tSNE, which shift clusters randomly from

run to run based on initialization, is solved by Multiscale PHATE, which can faithfully create the same embedding across multiple runs with different initializations (Extended Data Figure 1d).

### **Selection of clusters based on MELD mortality likelihood score to infer clinical associations**

Multiscale PHATE abstracted data is amenable to many downstream computational analysis tools like MELD [15] for comparative analysis and Density Resampled Estimate of Mutual Information (DREMI) [16] for computing mutual information between markers within subpopulations (Figure 1d). In particular, to identify populations that are differentially enriched between experimental conditions or patients, we combine our powerful multigranular clustering approach with MELD [15] (Extended Data Figure 1b). MELD creates a joint graph of the samples being compared, and returns a relative likelihood that quantifies the probability that each cell state in the graph is more likely in the control condition (which corresponds here to patients with positive outcome) or experimental condition (which corresponds here to patients with adverse outcomes). This likelihood score highlights regions of the manifold enriched in different conditions.

Finding a clustering method that matches the level of granularity of relative likelihood is a difficult problem, requiring the computationally complex vertex frequency clustering solution proposed previously [15]. However, Multiscale PHATE offers an alternative, less computationally expensive solution as one of the granularities identified by diffusion condensation matches the clusters revealed by the MELD likelihood score. Combining the likelihood signal with our multigranular analysis, we can identify populations that are associated with particular outcomes with greater accuracy than other methods (Extended Data Figure 3e).

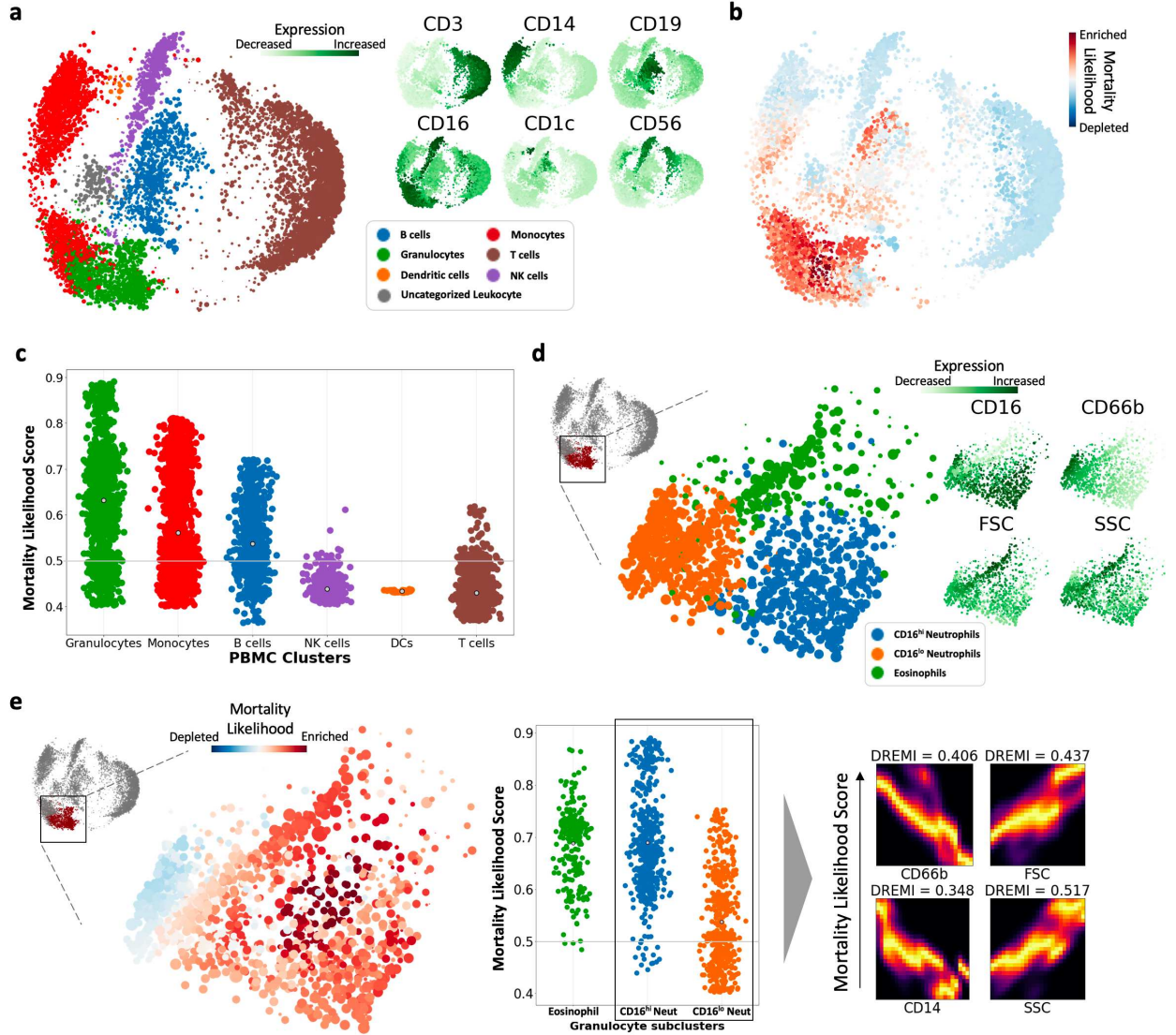
### **Construction of patient manifold through multiresolution cluster evaluation**

After creating a cellular manifold by integrating hundreds of patients samples, it is critical to understand how similar or different each of these patients are from one another. Uncovering sample level density variations along the cellular manifold can be a powerful strategy to identify patient clinical states that are similar or dissimilar from one another. With the goal of creating a manifold of patients, where each point represents a unique patient sample and distances between points represent how similar or different the underlying samples are in their cellular states as measured by flow cytometry, we evaluate clusters at multiple levels of the condensation topology. First, we identify all clusters at a particular level of the condensation topology. For a particular patient sample, we identify the proportion of the sample’s cells that fall within each of these clusters, creating a vector of cell proportions. This process is first applied to a single resolution, creating a set of features for every sample within this resolution, and is then repeated for all samples across multiple resolutions to create an even richer, multiscale set of features related to prior work in multiresolution optimal transport [23]. This high dimensional multiscale feature matrix can then be embedded with PHATE for visualization.

### **3.3 Multiscale PHATE analysis of 251 SARS-CoV-2 patient blood samples reveals subsets of cells associated with mortality:**

One hundred sixty eight patients with moderate to severe COVID-19 [24] were admitted to YNHH and recruited to the Yale IMPACT (Implementing Medical and Public Health Action Against Coronavirus CT) study. From each patient, blood samples were collected across multiple timepoints to characterize patient cellular responses across the spectrum of disease. In total, the composition of peripheral blood mononuclear cell (PBMC) was measured by flow cytometry on 251 samples.

Finally, clinical data was extracted from the electronic health record corresponding to each biosample timepoint to allow for clinical correlation of findings (see Methods). In this analysis, we define poor or adverse outcomes as patients who died from infection, while good outcomes as patients who survived. In order to analyze over 54 million cells characterized across 4 different sets of flow marker panels, we applied Multiscale PHATE to identify subsets of peripheral blood mononuclear cells (PBMCs) associated with mortality and survival.



**Figure 3:  $CD16^{hi} CD66b^{lo}$  Neutrophil subset enriched in patients who die from COVID-19.**  
a. Multiscale PHATE visualization of PBMCs identifies all major cell types based on cell type specific markers.  
b. Visualization of mortality likelihood score computed by MELD.  
c. Visualization of mortality likelihood score organized by cell type reveals enrichment of granulocytes, monocytes and B cells in patients who die from COVID-19.  
d. Zoom in of granulocyte population identifies subsets of neutrophils and eosinophils based on expression of known markers.  
e. Visualization of mortality likelihood score in granulocyte population identifies  $CD16^{hi}$  neutrophils enriched in patients with worse outcomes. Key associations between markers and mortality likelihood scores in neutrophils computed by DREMI and visualized with DREVI.



## Key dysfunctional myeloid, granulocyte and B cell subsets are enriched in patient who die from infection:

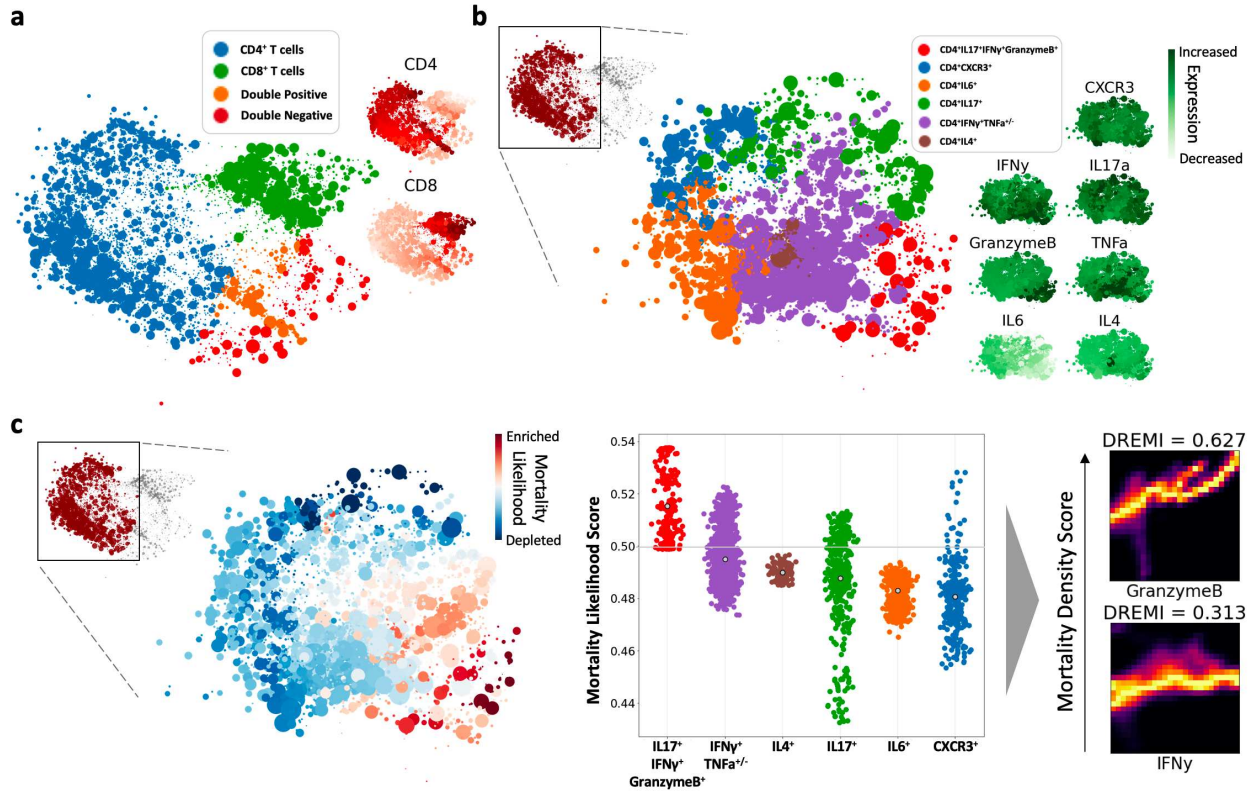
To explore the role of individual PBMC cell types in disease pathogenesis, we examined 22 million cells measured on a myeloid-centric flow cytometry panel from 210 patient samples suffering from COVID-19. Using cell type specific marker staining, we characterized Multiscale PHATE clusters (Figure 3a). Using MELD and the mortality outcome for each patient in our cellular state space, we were able to compute the mortality likelihood score, which identified cellular states enriched in patients who die from infection (darker red) or patients that survive (darker blue) (Figure 3b). When mapping these scores onto cluster labels, we found that the three populations most enriched in mortality were granulocytes ( $CD16^+SSC^{hi}$ ), B cells ( $CD19^+$ ), and monocytes ( $CD14^+$ ) while the population most enriched in survival was T cells ( $CD3^+$ ) (Figure 3c).

**Resting population of circulating neutrophils enriched in with patients who die from COVID-19.** We zoomed in on the granulocyte population and identified  $CD16^{hi}$  neutrophils,  $CD16^{lo}$  neutrophils and eosinophils based on the expression of CD16, CD66b, granularity by side scatter (SSC) and size by forward scatter (FSC) (Figure 3d). After mapping our mortality scores onto this granulocyte population we found that the  $CD16^{hi}$  neutrophils were enriched in patients who died from infection. In order to identify which cellular markers beyond CD16 were most correlated with mortality in neutrophils, we computed DREMI between protein expression and mortality likelihood scores in both neutrophil subsets. We identified that while CD14 and CD66b were negatively correlated with mortality, increased FSC and SSC were both strongly positively correlated with mortality in neutrophils, indicating that  $CD16^{hi}CD66b^{lo}$  neutrophils were enriched in patients that died from COVID-19 (Figure 3e). Based on the PBMC isolation protocol used (see Methods) neutrophils obtained were by definition Low-Density Neutrophils, containing both the mature and immature subsets. Considering the sensitivity of CD16 expression, high CD16 in our cohort was most likely indicative of a mature population that has not responded to an activating stimulus [25–27]. Neutrophils from patients with worse disease also expressed less CD66b; in contrast, an increase in surface expression of CD66b occurs following degranulation [28]. The combination of high complexity, high CD16 expression, and low CD66b expression suggests a resting population of circulating neutrophils present in patients with lethal disease.

**$CD14^-CD16^{hi}HLA-DR^{lo}$  monocyte subset associated with patient mortality.** In order to identify monocyte subsets implicated in disease, we zoomed in to the monocyte population and identified major subtypes based on the expression of markers CD16 and CD14 (Extended Data Figure 4a). The combination of these markers allowed us to distinguish between  $CD14^+CD16^-$  monocytes,  $CD14^+CD16^{int}$  monocytes and  $CD14^-CD16^{hi}$  monocytes. After mapping our computed mortality likelihood scores onto this population, we identified that  $CD14^-CD16^{hi}$  monocytes were the most strongly enriched in severe infection, followed by  $CD14^+CD16^{int}$  monocytes (Extended Data Figure 4b). These findings agreed with published observations as others have also noted an influx of  $CD14^+CD16^{int}$  and  $CD14^-CD16^{hi}$  monocytes in the lungs of patients with severe disease [11, 29, 30]. Furthermore, across all monocytes, CD16 was positively correlated with mortality while CD14 and HLA-DR were correlated with survival, identifying a distinct  $CD14^-CD16^{hi}HLA-DR^{lo}$  population of monocytes enriched in mortality. The loss of HLA-DR has been previously shown in monocytes from COVID-19 patients [31]. Monocytes expressing HLA-DR can serve as antigen-presenting cells to shape the adaptive T cell response, but monocytes in this cohort, expressing reduced amounts of HLA-DR, would likely have very limited capacity to prime effector T cell responses. Interestingly, a similar phenomenon occurs in sepsis patients, as well, and is indicative of worse outcomes [32–34].

In this setting, elevated levels of IL-10 have been linked to decreased HLA-DR on monocytes [35,36]. As COVID-19 patients also present with significantly elevated levels of IL-10 in circulation, a similar mechanism may be at play here [31,37].

**Multiscale PHATE identified plasmablast population associated with mortality.** There has been a persistent interest in the role of B cells during disease due to their potential to generate neutralizing antibodies. In our broad PBMC analysis, however, B cells were among the most enriched populations in severe outcomes (Figure 3c). In order to explore B cells in greater detail, we processed 154 patient samples on a B cell specific flow cytometry marker panel. Analyzing these cells by Multiscale PHATE granted us an unbiased, granular look at B cell subsets which would otherwise be difficult by traditional two-dimensional gating, popular for flow cytometry analysis. These subsets include transitional B cells ( $\text{IgD}^+\text{IgM}^+\text{CD27}^-/\text{CD38}^+\text{CD24}^+$ ), naïve B cells ( $\text{IgD}^+\text{IgM}^+\text{CD27}^-/\text{CD38}^-$ ), switched ( $\text{IgD}^-\text{IgM}^-\text{CD27}^+$ ) and unswitched memory B cells ( $\text{IgD}^+\text{IgM}^+\text{CD27}^+$ ), activated B cells ( $\text{IgD}^-\text{CD138}^-\text{CD86}^+\text{HLADR}^+$ ), and antibody secreting cells ( $\text{CD138}^+\text{CD38}^+$ ) (Extended Data Figure 4c). After identifying these major cell types, we computed mortality likelihood scores to identify B cell subtypes implicated in mortality. Interestingly, the most enriched cell type in patients with adverse outcomes was a subset of the antibody secreting population defined by  $\text{CD86}^{\text{lo}}\text{HLADR}^-/\text{CXCR3}^+$ , also known as plasmablasts. Meanwhile the cell types most enriched in patients with good outcomes was a subset of late activated mature B cells defined by  $\text{CD86}^+$  (Extended Data Figure 4d). Despite the well-described protective roles of circulating antibodies, these results are consistent with earlier findings from COVID-19 patients, which discuss the potential role of B cells in disease pathogenesis [38–40]. Given the abundance of circulating IL-6 in COVID-19 patients, it is possible that in this setting, IL-6 skews B cell differentiation into antibody secreting cells [41]. Considering the lack of mutations in neutralizing, anti-SARS-CoV-2 antibodies, skewing toward antibody secreting cells may come at the expense of transit through the germinal center, thus producing a less potent or non-specific antibody response [42].



**Figure 4: Multiscale PHATE identifies Th17 subset enriched in patients who die from COVID-19**

a. Multiscale PHATE visualization of T cell focused cytokine panel identifies broad T cell subtypes.  
b. Zoom in of  $CD4^+$  T helper cells identifies subsets based on expression of functional markers.  
c. Visualization of mortality likelihood score identifies  $IFN\gamma^+$  GranzymeB $^+$  Th17 cell enrichment in patients with poor outcomes. Key associations between markers and mortality likelihood scores are computed by DREMI and visualized with DREVI.

### Key pathogenic T cell subsets are enriched in patients who die from infection:

Although T cells collectively were enriched in patients who recovered from infection (Figure 3c), there are a diverse set of T cell subsets which have been implicated in severe disease pathogenesis. In order to identify functional T cell subsets enriched in patients who died from COVID-19, we applied Multiscale PHATE to 22 million T cells measured on a cytokine-specific flow cytometry panel. After identifying salient levels of granularity for downstream analysis, we identified both  $CD4^+$  and  $CD8^+$  T cell subsets at coarse granularity (Figure 4a).

**Fine grain analysis of protective T cell population helps identify pathogenic  $IFN\gamma^+$  GranzymeB $^+$  Th17 subpopulation.** Using Multiscale PHATE's zoom and cluster capabilities, we were able to visualize the  $CD4^+$  T cells and subdivide the cells into functional subsets using functional markers,  $IFN\gamma$ , IL-17, and IL-4 (Figure 4b). Interestingly, in our dataset, we identified two different subsets of  $CD4^+$  IL-17 producing T cells classically known as Th17 cells, one co-producing GranzymeB and  $IFN\gamma$  and one staining negative for both markers. Finally, we classify two final subsets based on the expression of IL6 and CXCR3 (Figure 4b). To identify cell types enriched in mortality, we computed a mortality likelihood score. By organizing our scores by T helper subset, it became clear that the Th17 subset co-producing  $IFN\gamma^+$  GranzymeB $^+$  were enriched in patients who



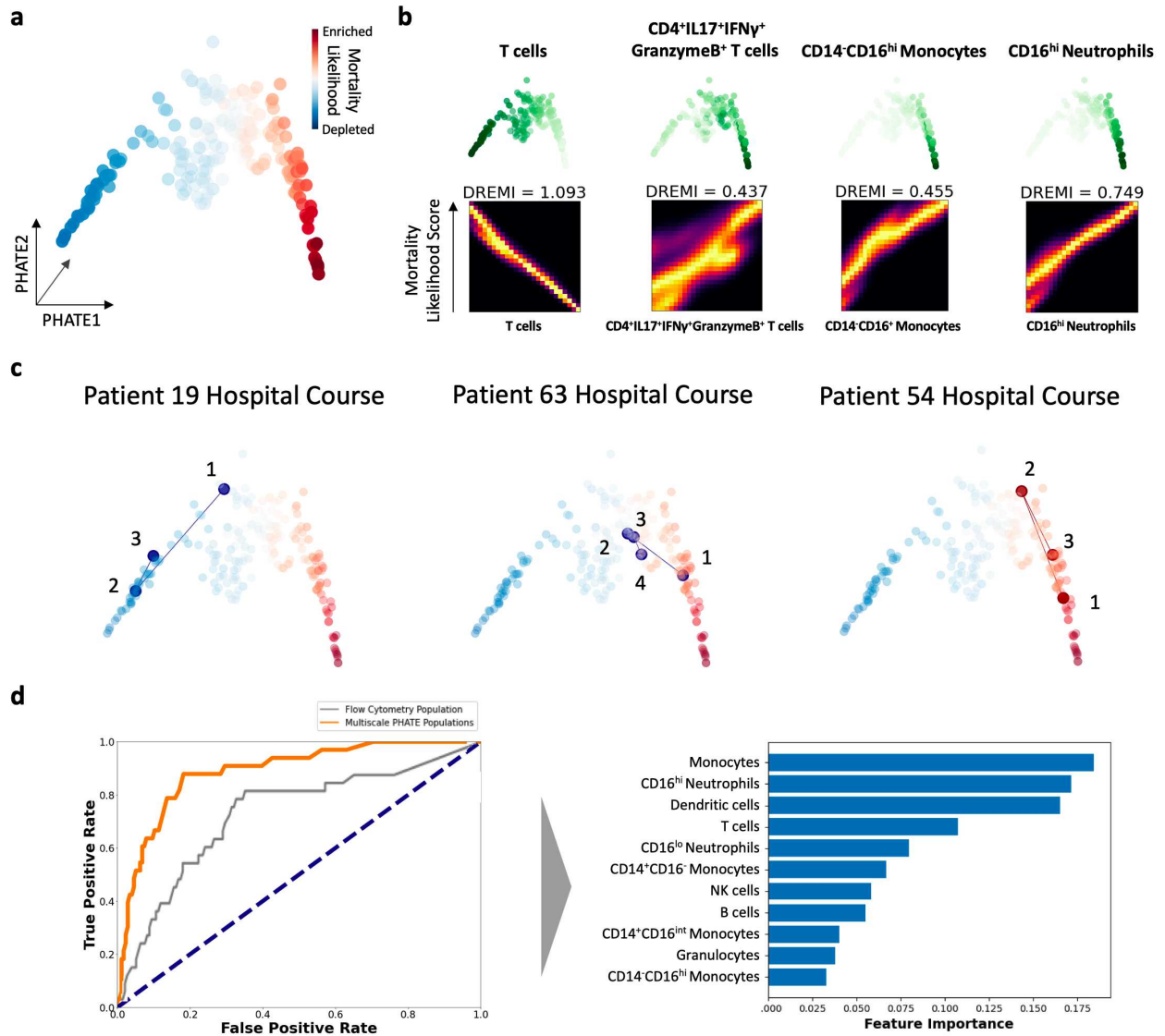
died from infection. Furthermore, GranzymeB and IFN $\gamma$  were positively associated with mortality likelihood on DREMI analysis across all CD4 $^{+}$  T cell subsets (Figure 4c).

While Th17 cells can play protective roles [43], IFN $\gamma^{+}$ GranzymeB $^{+}$  Th17 cells have been associated with tissue damage. In a model of murine autoimmune encephalomyelitis, a discrete subset of IFN $\gamma^{+}$ GranzymeB $^{+}$  Th17 cells caused significantly worse disease than traditionally activated Th17 cells [44]. Previous literature had also observed that high levels of circulating IL-17 produced from IFN $\gamma^{+}$ GranzymeB $^{+}$  Th17 cells could drive a strong pro-inflammatory immune response and promote neutrophil expansion. Likewise, recent reports have indicated the harmful contribution of neutrophils and neutrophil extracellular traps (NETs) in SARS-CoV-2 infections [45–47]. This influx of neutrophils can be further exacerbated by the virus-induced loss of ACE2 [48], and cumulatively, these events have the potential to trigger ARDS, as seen in COVID-19 patents. However, what regulates neutrophil recruitment, survival, and subsequent NET release during the disease has not been definitively identified in COVID-19. Interestingly, patients with adverse outcomes in this cohort demonstrated an enrichment in IFN $\gamma^{+}$ GranzymeB $^{+}$  Th17 cells, as well as CD16 $^{+}$  neutrophils. We posit that IFN $\gamma^{+}$ GranzymeB $^{+}$  Th17 cells in our cohort may precipitate these pathogenic effects via IL-17 secretion. While our flow cytometry data was limited to identifying cells in circulation, sequencing data from upper respiratory tracts of COVID-19 patients observed Th17 cells in the airways, as well [49,50]. Either in the lungs or in circulation, IFN $\gamma^{+}$ GranzymeB $^{+}$  Th17 may influence neutrophil activity by inducing IL-8 release from airway epithelial cells or G-CSF from microvascular pericytes [51–53]. It was recently shown that MAIT cells comprise a substantial portion of IL-17 expressing cells in the upper respiratory tracts of COVID-19 patients; consequently, secretion of IL-17 in the lung may not be primarily confined to the Th17 compartment. The two may also act synergistically as MAIT cells have been shown to promote the recruitment of activated CD4 $^{+}$  T cells to the lungs during pulmonary infection [54].

**Hyperactivated CD8 $^{+}$ TIM3 $^{+}$ HLA-DR $^{+}$ PD1 $^{+}$  TEMRA cells, expressing GranzymeB enriched in patients who die from COVID-19.** In acute viral infections, CD8 $^{+}$  T lymphocytes play a critical role in the clearance of virus [55]. By the directed secretion of Granzyme B, these effectors may rapidly kill virally-infected targets [56]. In order to characterize the role of CD8 $^{+}$  T cell subsets in disease, we zoomed in on CD8 $^{+}$  T cells in our cytokine-focused T cell panel. Using the expression of cell surface markers and cytokines, we identified three major subsets, one producing GranzymeB, one producing IFN $\gamma$  and one producing TNF $\alpha$  (Extended Data Figure 5a). After mapping mortality likelihood scores onto the CD8 $^{+}$  subpopulation, it became clear that the GranzymeB $^{+}$  population is most enriched in mortality with GranzymeB expression being highly associated with mortality in CD8 $^{+}$  T cells (Extended Data Figure 5b). These findings are consistent with a previous study of SARS-CoV-2 infected patients that observed an association between the enrichment of CD8 $^{+}$  T cells expressing high amounts of GranzymeB and increased disease severity [57]. Despite the protective role of GranzymeB in other viral infections, our data and others indicate that its excess may lead to worse outcomes, including mortality. This possibility is supported by early findings of GranzymeB $^{+}$  CD8 $^{+}$ -induced tissue damage in different murine models of respiratory viral infections [58,59] or from clinical studies [60]. In these early studies from mice, pathogenic GranzymeB $^{+}$  CD8 $^{+}$  T cell activity manifested in the presence of extremely high viral loads or in the absence of other lymphocytes and antigen-specific antibodies. Likewise, all of these factors are present in COVID-19 patients- high viral loads, lymphocytopenia, and ineffective antibody responses - which permits the emergence of this hyper-activated CD8 $^{+}$  population associated with more severe disease. To gain additional insight on which discrete subset of CD8 $^{+}$  T cells may be the source of GranzymeB, we performed detailed surface staining of all T cells.

We analyzed 208 patient samples using a flow cytometry panel containing markers indicative of T cell subset identity and activation status. After identifying the ideal granularity to analyze the data, we identified CD4<sup>+</sup>, CD8<sup>+</sup> and double positive T cell subsets (Extended Data Figure 5c). Zooming in to the CD8<sup>+</sup> subset, we identified a range of activation states based on the expression of key markers: Effector (TIM3<sup>+</sup>PD1<sup>+</sup>/CD45RA<sup>+</sup>), Follicular (CD45RA<sup>-</sup>/CCR7<sup>-</sup>CD127<sup>-</sup>/CXCR5<sup>+</sup>PD1<sup>+</sup>), Memory (CD45RA<sup>-</sup>/CCR7<sup>+</sup> or CCR7<sup>-</sup>CD127<sup>+</sup>), Naive (CD45RA<sup>+</sup>/CCR7<sup>+</sup>CD127<sup>+</sup>) and T effector memory cells re-expressing CD45RA (TEMRA) (CD45RA<sup>+</sup>/CCR7<sup>-</sup>CD127<sup>lo</sup>) (Extended Data Figure 5d). After computing MELD mortality likelihood score, we identified that the TEMRA population displayed the most enrichment in severe infection. Furthermore, across all CD8<sup>+</sup> T cells, activation state markers PD1, TIM3, HLA-DR and CD45RA were also positively correlated with mortality on DREMI analysis, while markers like CD127 and CCR7 were negatively correlated with mortality (Extended Data Figure 5e). Our findings here are in agreement with contemporaneous studies of SARS-CoV-2 patients [38, 57, 61]. Cumulatively, our data correlate mortality with a hyper-activated CD8<sup>+</sup> T cell response in the form of CD8<sup>+</sup>CD45RA<sup>+</sup>TIM3<sup>+</sup>HLA-DR<sup>+</sup>PD1<sup>+</sup> TEMRA cells, likely expressing GranzymeB.

### 3.4 Multiscale patient manifold construction reveals potential mechanisms of disease:



**Figure 5: Patient manifold corroborates cellular states associated with disease pathogenesis.**  
*a. Visualization of patient manifold via PHATE and mortality likelihood score based on patient outcomes computed via MELD.*  
*b. Visualization of key cell population enrichment trends over the manifold with associations computed by DREMI and visualized with DREVI.*  
*c. Tracing three patients' hospital courses over patient manifold. Patients 19 and 63 were discharged while patient 54 died.*  
*d. Comparing predictability of patient mortality using random forest classifier on Multiscale PHATE identified populations and flow cytometry identified populations. Most predictive Multiscale PHATE clusters are ranked through feature importance analysis.*

Here, we show that Multiscale PHATE-derived clusters across multiple scales form a rich set of feature descriptors for patients measured in single cell modalities. Although, the purpose of measuring single cell data is indeed to derive features in the form of cells, patients can be hard to compare and

analyze at this level. Common approaches simply compare cluster proportion or averaged expression levels across patients. However, since Multiscale PHATE creates cellular groupings at multiple granularities, we can derive a rich summarization of patients across scales.

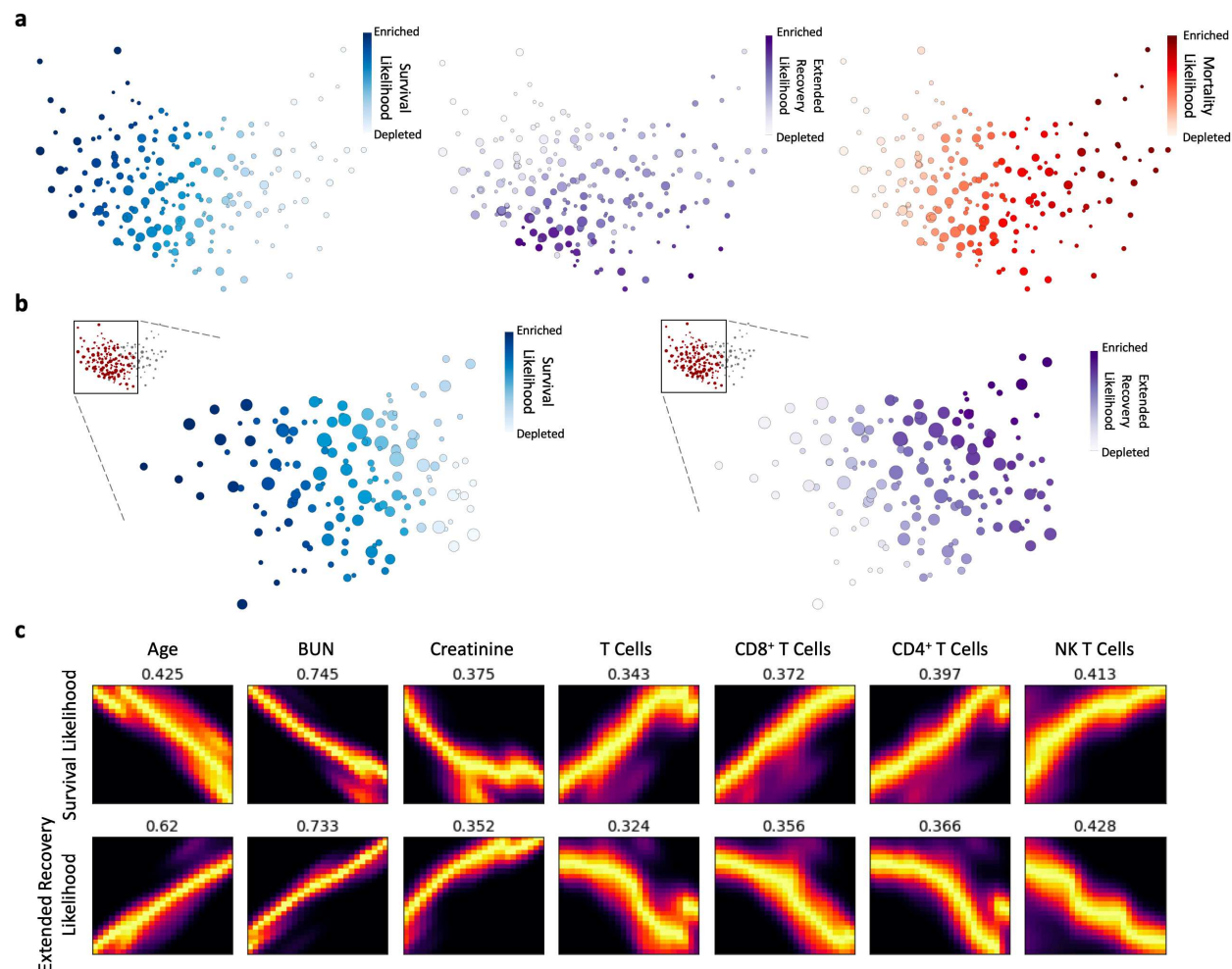
We create a patient-embedding using cluster proportions from several levels of the condensation topology of the myeloid-focused flow cytometry using our patient manifold approach (Figure 5a). Briefly, the proportion of a patient’s cells that belong to clusters at several levels of the topology are used as a feature vector (see Methods). These patient descriptors are then embedded and visualized with PHATE [13]. The resultant embedding demonstrates that the patients lie on a low dimensional continuum or manifold themselves. When the patient embedding is colored by the manifold-based likelihood estimate of mortality outcomes, we see that the dominant progression in the data is indeed clinical outcome, with patients on the left enriched for good outcomes (darker blue) and patients on the right enriched for adverse outcomes (darker red).

In order to associate previously identified cellular populations with outcome, we computed DREMI between these population proportions and mortality likelihood score. We identified that while T cells overall were negatively correlated with mortality,  $CD4^+IFN\gamma^+GranzymeB^+Th17$  cell, plasmablasts,  $CD16^{hi}$  neutrophils and  $CD14^-CD16^{hi}$  monocytes were all strongly positively associated with mortality (Figure 5b). These findings indicate that precipitous decline in T cells correlates with mortality, while subsets of neutrophils, monocytes and Th17 cells, all previously highlighted in our analyses, are increased in patients with adverse outcomes. Finally, we trace clinical states of three patients, 19, 63 and 54, across the patient manifold to determine if our construct accurately recapitulates patient trajectories. Surviving patients 19 and 63 had their clinical trajectories consistently go from the high mortality region to the low mortality region. In contrast, patient 54, who succumbed to disease, had a tortuous set of clinical states all of which mapped within the high mortality region (Figure 5c).

To identify if age, sex and other clinical variables were preferentially associated with mortality on our construction, we mapped these clinical variables onto the patient manifold. We found that patients who were more likely to experience poor outcomes were also more likely to be older, male, receive ventilatory support and have higher markers of inflammation (Extended Data Figure 6a). We ran DREMI analysis to find associations between these clinical variables and key cell types implicated in infection pathogenesis. We found that females and young individuals were more likely to mount a robust T cell response. This finding builds upon a body of literature that shows immune responses may differ across sex and age [62, 63]. Specifically, for SARS-CoV-2, this finding corroborates a separate analysis that found that activated T cells play a protective role in women but not as much in men [64]. Additionally, the negative relationship between age and T cell numbers has been extensively studied [65, 66]. Our analysis finds the same trend in our cohort of patients, who are in general older, predisposing them to requiring hospitalization. Epidemiological data analyzing large numbers of COVID-19 patients enumerate the significant contributions of age and sex for disease severity [67, 68].

In order to see if Multiscale PHATE-derived sub-populations could predict disease outcome, we combined the features of patients that we identified in our myeloid-focused flow cytometry panel with clinical outcome to train a random forest classifier (see Methods). Using these abstracted features, we accurately predict outcome in 83.5% of cases. When performing a similar prediction task using flow cytometry gated populations, we were only able to predict outcome with a lower accuracy of 73.8%. Furthermore, we identified that monocytes,  $CD16^{hi}$  neutrophils and T cells were three of the top four cell types most predictive of eventual disease outcome in our classifier model (Figure 5d).

### 3.5 Multiscale clinical manifold construction highlights potential mechanisms of disease convalescence:



**Figure 6: Multiscale manifold of patient clinical features identifies cell types associated with extended COVID-19 recovery phase**

a. Visualization of Multiscale PHATE clinical manifold constructed on patient clinical features. Embedding is colored by likelihood scores based on patient outcomes computed via MELD.

b. Zoom in on transition point between high extended recovery likelihood score and high survival likelihood score.

c. Patient clinical features and flow cytometry identified cell populations associated with patient outcomes using DREMI and visualized with DREVI.

Thus far, we have primarily used Multiscale PHATE to identify multiresolution structure in single cell flow cytometry data. We now showcase the utility of Multiscale PHATE on a different type of data: laboratory, clinical, and demographic data generated from routine clinical care of COVID-19 patients admitted to YNHH. Patients admitted for severe disease are characterized largely by having advanced age and have systemic infection leading to multiorgan dysfunction. Using 18 clinical and demographic measurements collected on 2,135 patients admitted to YNHH diagnosed with COVID-19, we created a multiscale embedding capturing patient states across the spectrum of disease severity. Patient outcomes at discharge were categorized as discharge to home, discharge



to rehabilitation for extended recovery, and discharge to hospice or death while in hospital. Using each of these outcomes, we computed likelihood scores with MELD corresponding to each outcome: survival likelihood score, extended recovery likelihood score and mortality likelihood score (Figure 6a).

In order to understand how clinical features could inform outcomes, we computed DREMI and DREVI analysis between clinical features and each of our likelihood scores (Extended Data Figure 7a-b). As anticipated, markers of physiologic instability, such as decreased systolic blood pressure and increased respiratory rate, as well as systemic inflammatory markers, increased ferritin, procalcitonin and CRP, were associated with higher mortality. Beyond these general markers, however, several markers of organ dysfunction were also strongly associated with mortality. Specifically, kidney dysfunction, as measured by blood urea nitrogen (BUN) and creatinine, as well as liver dysfunction, aspartate aminotransferase (AST) and alanine aminotransferase (ALT), were correlated with mortality. Although COVID-19 most commonly involves the respiratory system, these findings are consistent with clinical reports of severe disease from a generalized inflammatory state resulting in multiorgan damage and failure.

A subset of patients infected with SARS-CoV-2 experience prolonged recovery periods. In fact, our multiscale embedding of patient clinical states suggests a transition between a region of high survival likelihood score and a region of high extended recovery likelihood score (Figure 6a). In order to understand which cellular populations and clinical features drive the difference between these outcomes, we decided to zoom into this transition point (Figure 6b). We computed DREMI association scores between clinical features and flow sorted cellular populations to identify features differentially associated with survival and extended recovery. Our analysis found that age and kidney dysfunction were strongly associated with extended recovery indicating that older patients with worse kidney function were more likely to experience lengthy recovery periods from infection (Figure 6c).

Beyond clinical features like age and organ function, we also discovered different cellular populations associated with outcomes such as survival or long-term recovery from disease (Figure 6c), thus showcasing the way in which Multiscale PHATE can be used as a substrate for integrating and analyzing multiple modalities of data. Multiple recent publications have addressed the question of protection conferred by T cell-mediated responses [69–71] - with many demonstrating that antigen-specificity and T cell responsiveness lead to improved outcomes [72, 73]. Our analyses are in line with these findings, indicating that though some T cell subsets may be pathogenic, the major T cell subsets overall are positively associated with survival and negatively associated with a lengthy recovery phase. Interestingly, no myeloid subsets were found to be associated with length recovery periods, indicating that T cells and T cell subsets are perhaps more associated with recovery length while other immune populations may be associated with mortality.

## 4 Discussion

Here we presented a multiscale data exploration technique to visualize, understand and compare large-scale datasets, filling a key gap in biological data exploration. Multiscale PHATE finds groupings of data at varying scales that are predictive of clinical outcome. Biological data naturally contains multi-granular structure. Most analysis methods, however, whether clustering or dimensionality reduction algorithms, generally only look at a single level of resolution and do not offer a systematic way to explore different scales. Hierarchical clustering is one method that could offer certain scales of resolution. However, due to the constant merges that occur in hierarchical clustering approaches, like Louvain, many levels of resolution are missed and biologically important levels of granularity are

not recapitulated. By contrast, Multiscale PHATE offers a fast manifold learning-based technique for uncovering a continuum of resolutions of structure and features by understanding data topology. The speed and effectiveness of Multiscale PHATE is due in large part to key algorithmic advances we presented to make the underlying diffusion condensation process scalable, and representational advances in using diffusion potential coordinates as the substrate for the condensation.

We show that Multiscale PHATE can be combined with other techniques, like manifold density estimation (MELD), mutual information (DREMI), and classification (random forest classification) to provide deep and detailed insights in biological processes. We showed several effective examples of combining Multiscale PHATE with a technique known as MELD that shows the relative likelihood of seeing cells from different categories of patients in different parts of the cellular manifold. When MELD is combined with Multiscale PHATE, we can find levels of resolution that naturally capture the salient differences between patients with different clinical outcomes. Interestingly, Multiscale PHATE’s ability to zoom in helps identify pathogenic subsets of protective populations. Across our analyses, T cells have been shown to be protective against poor outcomes, corroborating previous work done in COVID-19. While broadly this cell type is protective, a multiscale zoom in of  $CD4^+$  T cells reveals a pathogenic  $CD4^+IFN\gamma^+GranzymeB^+Th17$  subpopulation. The multiresolution analysis we performed stresses the need to analyze data at multiple granularities. While broad cell types, such as T cells, may appear to be protective, smaller cellular subsets, such as pathogenic Th17 cells, may actually be driving patient mortality. In our work, we show several instances where DREMI between MELD likelihood scores and Multiscale PHATE identified clusters revealed potential associations between outcomes and key subpopulations, like  $CD16^+CD14^-CD66b^-$  neutrophils and  $CD14^{lo}CD16^+HLA-DR^{lo}$  monocytes. Furthermore, we showed that Multiscale PHATE identified populations combined with outcome variables can be used to predict clinical outcomes better than the current gold standard for flow cytometry analysis. Finally, we show that our approach is generalizable to a wide variety of biomedical data, including scRNAseq, scATACseq, CyTOF, TCR repertoire sequencing and clinical datasets.

While we have demonstrated Multiscale PHATE in the context of COVID-19 patient data, we believe that both the technique and the ways in which we have used it to analyze multi-modal data are widely applicable. Multiscale PHATE can also be used with an individual data modality to uncover structure where canonical cellular subtypes are not available, such as in patient-specific cancer or tumor cell types. Generally, as datasets continue to increase in size and the number of samples continue to expand, our scalable algorithm will become even more critical for joint analysis.

## 5 Methods

### 5.1 Computational Methods

In the following sections we provide a thorough description of each aspect of the Multiscale PHATE algorithm and the use of downstream analysis tools. This includes but is not limited to explanations of algorithm design choices, information on how comparisons between algorithms were run and details on how the patient manifold was constructed.

#### 5.1.1 Diffusion information geometry for visualization and condensation

The multiresolution visualization provided by Multiscale PHATE relies on the construction of a diffusion geometry that captures the intrinsic structure of the data. Such a construction was first presented in the context of manifold learning with Diffusion Maps (DM), which rely on diffusion coordinates derived from spectral decomposition of the heat kernel over (Riemannian) manifolds [74].



The DM construction approximates the heat kernel on data by defining a Markovian diffusion process whose transition probabilities are given by  $p(x, y) = \frac{k(x, y)}{\|k(x, \cdot)\|_1}$ , where the  $L_1$  norm is taken over the input data and  $k(\cdot, \cdot)$  is a kernel function for capturing the similarity between local neighborhoods in the data. Then, an integral diffusion operator is constructed as  $\mathbf{P}f(x) = \int p(x, y)f(y)dy$ , which is represented in finite settings as a matrix with entries  $[\mathbf{P}]_{ij} = p(x_i, x_j)$ , where  $\{x_1, x_2, \dots\}$  are the input data points (e.g., cells or strains in our case). By taking powers of this diffusion operator, we can consider  $t$ -step diffusion probabilities between data points given by  $p^t(x_i, x_j) := \Pr[x_i \xrightarrow{t\text{-steps}} x_j] = [\mathbf{P}^t]_{ij}$ . Finally, the diffusion geometry considers each data point  $x$  via its  $t$ -step diffusion distribution  $p_x^t = p^t(x, \cdot)$ , and DM aims to extract low dimensional coordinates where Euclidean distances capture a diffusion distance metric defined as  $L_2$  distances between these distributions, called *diffusion distances*.

While several kernels are used in practice to construct the diffusion operator  $\mathbf{P}$ , a standard choice is the Gaussian affinity  $k_\varepsilon(x, y) = \exp(-\|x - y\|^2/\varepsilon)$  [13, 74–76], in which case we denote the diffusion operator  $\mathbf{P}_\varepsilon$ , where  $\varepsilon$  determines the neighborhood radius. This kernel choice is often seen in theoretical and mathematical work due to its established properties on data sampled from locally low dimensional geometries (i.e., data manifolds) [74, 77, 78]. In particular, it can be verified that when the data is sampled from a Riemannian manifold, the diffusion operator  $\mathbf{P}_\varepsilon^{t/\varepsilon}$  constructed from  $k_\varepsilon(\cdot, \cdot)$  converges to the heat kernel on the underlying manifold as  $\varepsilon \rightarrow 0$ . Further, as  $\varepsilon \rightarrow 0$ , the eigenvectors of  $\mathbf{P}_\varepsilon^{t/\varepsilon}$  operator converge to eigenvectors of the Laplace-Beltrami operator that characterize the solutions of the heat equation ( $\partial_t f(x, t) = \nabla_x^2 f(x, t)$ ) with von Neumann boundary conditions on the underlying manifold of the data. Based on these convergence properties, the embedding provided by DM requires an eigendecomposition of  $\mathbf{P}^t$  to its eigenvalues  $1 = \lambda_0^t \geq \lambda_1^t \geq \lambda_2^t \geq \dots \geq 0$  and corresponding eigenvectors  $\phi_0, \phi_1, \phi_2, \dots$ , which then yield the diffusion coordinates  $x \mapsto (\lambda_j^t \phi_j)_{j=1}^\eta$ , where  $\eta$  is determined by the numerical rank of  $\mathbf{P}^t$  and the first eigenpair is typically discarded since  $\phi_0$  is always constant. We refer the reader to [74] for more details on DM and its properties.

While the analytic relation between spectral embedding with diffusion coordinates is appealing from a manifold learning perspective, the resulting DM often separates trajectories, pathways, or clusters into independent eigenspaces. This, in turn, yields multidimensional representations that cannot be conveniently visualized (e.g., having significantly more than 2-3 dimensions), and more importantly, cannot be directly projected into 2D or 3D displays that faithfully capture diffusion distances. In order to overcome this and extract a low dimensional data visualization, the recently proposed PHATE method treats the constructed diffusion geometry as a statistical manifold and leverages tools from information geometry to define a family of diffusion information distances defined as  $D_t^\gamma(x, y) = \left\| \Delta_{(x, y)}^{(\gamma)}(\cdot) \right\|_2$  where

$$\Delta_{(x, y)}^{(\gamma)}(z) = - \int_{p_x^t(z)}^{p_y^t(z)} u^{-\frac{\gamma+1}{2}} du = \begin{cases} p_x^t(z) - p_y^t(z) & \gamma = -1 \\ \log p_x^t(z) - \log p_y^t(z) & \gamma = +1 \\ \frac{2}{1-\gamma} \left[ (p_x^t(z))^{\frac{1-\gamma}{2}} - (p_y^t(z))^{\frac{1-\gamma}{2}} \right] & \text{otherwise} \end{cases} \quad (1)$$

and the parameter  $-1 \leq \gamma \leq +1$  attenuates the influence of lower probability differences in the overall distance. On one extreme ( $\gamma = -1$ ), the resulting metric yields the traditional diffusion distance. When  $\gamma = 0$ , it yields an  $f$ -divergence corresponding to Hellinger distances between diffusion distributions. On the other extreme ( $\gamma = +1$ ), the resulting information distance yields an  $L_2$  distance between localized diffusion energy potentials given by  $U_x^t(\cdot) = \log p_x^t(z)$ , as discussed in [13]. There, as well as in other work [79], it has been shown that this potential distance is amenable to a low dimensional embedding that captures and visually accentuates emergent global and local structures in the data - most importantly, trajectories and transitions between stable clusters in it. Therefore,

the PHATE method, as well as its variation here, are based on embedding potential distances directly into two or three dimensional coordinates via a stress-minimizing optimization procedure provided by multidimensional scaling (MDS). In addition to the core utilization of diffusion information geometry, the PHATE algorithm also includes robust construction of the initial neighborhood kernel, automatic tuning of diffusion resolution, and efficient sampling for scalability purposes. For more details about these aspects of PHATE, we refer the reader to [13].

Multiscale PHATE not only uses PHATE for visualization of several chosen iterations of the condensation process (explained below), representing multiple scales of data coarse graining, but also as the potential coordinate system that learns geometry of the data.

### 5.1.2 Multiresolution via the diffusion condensation time-inhomogeneous Markov process

The diffusion geometry underlying PHATE is naturally multiscale, via the diffusion time parameter  $t$  that controls the resolution of information captured by the diffusion process. Indeed, as the diffusion time increases, the distributions  $p_x^t(\cdot)$  (or potentials  $U_x^t(\cdot)$ ) consider increasingly diffused energy that attenuates local differences until eventually as  $t \rightarrow \infty$  all these distributions converge to a unique equilibrium stationary distribution, since the process is ergodic. However, as discussed in [80, 81], this process often diffuses information too rapidly to enable multiresolution representation of varying intermediate scales of data geometry. Further, in [13], it was shown that the diffusion time scale admits an optimal time scale for visualization, which can be identified automatically by distinguishing between a rapid denoising phase and a slow decay from metastable to equilibrium diffusion states. We can use this property to automatically tune the diffusion time scale to transitioning point between these two phases. However, here we aim to extend this analysis to provide a full multiscale or multiresolution data geometry, and therefore we need to provide better control of the propagation of information by intrinsic diffusion over the data.

One of the first attempts at alleviating the rapid convergence to stationary distribution in multiscale DM was presented in [80], as part of a hierarchical construction of localized diffusion folders (LDF) using a localized diffusion process, which was further analyzed in [81]. The localized diffusion process there limited each instantiation of the diffusion random walks to only traverse between two “diffusion folders” (i.e., clusters), thus blocking global pathways that quickly diffuse to wide regions in the data. While this process was shown to be effective in some applications involving hierarchical clustering, it requires separate clustering steps and a priori determination of scales at which to pause the diffusion and cluster into LDFs. Furthermore, the pruning of the diffusion process there is computationally intensive, as each diffusion affinity (or transition probability) requires simulating or approximating a local diffusion process between two considered clusters. However, the principles posed by this approach clearly established the need for careful manipulation of the underlying Markov process of DM in order to truly enable multiscale representation learning via diffusion geometry, and by extension the diffusion information geometry used in PHATE.

Interestingly, topological data analysis naturally creates multiscale structure by combining geometric and topological perspectives into a single framework. While studying data geometry is useful in understanding the precise measurements between objects, topological analysis is useful in describing the relationships between objects. A hybrid perspective can be appealing in situations such as ours, where geometry and relationships between data points are both important.

Persistent homology denotes a set of algorithms from the field of computational topology. The express goal of persistent homology is to describe the geometry—or shape—of sampled data sets, i.e., point clouds, at multiple scales, or granularities. The common scenario of persistent homology involves the analysis of distance functions on point clouds to approximate the manifold from which

a dataset  $X$  has been sampled. To this end, one typically constructs a simplicial complex [82], i.e., a generalization of a graph, which can contain higher-dimensional structural elements called simplices, corresponding to subsets of the data points of a certain cardinality. Given a distance threshold  $\delta$ , one typically constructs a simplicial complex containing all simplices whose pairwise distances are less than or equal to  $\delta$ , i.e.,

$$\mathcal{V}_\delta(X) := \{\sigma \subseteq X \mid d(x_i, x_j) \leq \delta\}. \quad (2)$$

Any subset  $\sigma$  of cardinality  $k - 1$  is called a  $k$ -simplices, meaning that the 0-simplices are the vertices of  $\mathcal{V}_\delta(X)$ , the 1-simplices are the edges, and so on (a graph can therefore be seen as a simple 1-dimensional simplicial complex). Each  $k$ -simplex  $\sigma$  in  $\mathcal{V}_\delta(X)$ , with  $k \geq 1$ , can be assigned a weight  $w_\sigma$  based on the pairwise distances of its underlying vertices by setting  $w_\sigma := \max_{x_i, x_j \in \sigma} d(x_i, x_j)$ . The resulting simplicial complex  $\mathcal{V}_\delta(X)$  can be seen as a backbone of the dataset  $X$ , combining geometry and scale information—via  $\delta$ —with topological information. The weighted simplices can be used to describe topological features such as connected components (0D), cycles (1D), and voids (2D) in  $X$ . Persistent homology therefore is a mixture between strictly geometrical approaches, which focus only on points, and strictly topological ones, which focus only on connectivity without incorporating distance—scale—information. The utility of persistent homology becomes apparent in characterizing the evolution of topological features at multiple granularities, determined via  $\delta$ . For a sufficiently large  $\delta$ , all points are connected to each other, yielding a complete graph or, equivalently, a full simplex, whereas for  $\delta \approx 0$ , almost all simplices in the complex will be vertices, i.e., higher-dimensional structure will be largely absent. It turns out that topological features can be efficiently tracked over multiple scales or granularities, i.e., multiple values of  $\delta$ , so that each feature is assigned a *persistence*. This quantity indicates over which granularities a feature is present. For instance, suppose that  $X$  is a densely-sampled square; it will feature one connected component with a high persistence value, because all points are connected for small values of  $\delta$  and will not be connected to another set of points for larger values.

Inspired by this topological data analysis, a more recent approach towards multiresolution diffusion-based coarse graining was presented in [12]. Diffusion condensation relies on replacing the traditional time-homogeneous Markov process typically used in diffusion frameworks [13, 74] with an inhomogeneous process, following the theoretical analysis in [83] that established the mathematical viability of diffusion geometry construction of such processes. Unlike previous approaches, the coarse graining in [12] does not rely on a clustering & pruning approach. Instead, it proposes to base the intuition for the diffusion construction from heat propagation that rapidly spreads over the data based on connectivity, to a condensation process that alternates between slow gravitation (e.g., as drops of water slowly gravitate towards each other) and fast merging, concentrated regions collapse (e.g., as water drops merge together) to a single point, creating a topological understanding of a dataset by calculating the persistence of individual points. If we view the merges of diffusion condensation as a change in terms of the topology of the dataset, the alternation between these meta-stable and transient regimes also provides a diffusion-analogous notion of persistence used in topological data analysis, which in turn naturally gives rise to emergent stable resolutions for multiscale visualization and clustering.

## Advancements in multiresolution visualization and clustering via Multiscale PHATE

In its original form, the time-inhomogeneous diffusion condensation process does is not optimized for visualizing complex and non-linear biological manifolds. Thus, we have modified diffusion condensation in the following ways:

1. Transforming datapoints to a novel diffusion potential coordinate system to learn the data manifold,
2. Computing a *fast* diffusion condensation process that scales to millions of data points,
3. Identifying levels for visualization based on gradient analysis and creating a density aware visualization.

### 5.1.3 Condensation on potential coordinates

The computation of the diffusion condensation process in [12] only uses the diffusion operator  $\mathbf{P}$ , interpreted as a low-pass (smoothing) filter that can be applied to any dataset encoded in a points-by-features data matrix  $\mathbf{X}$ . However, condensing in this feature space can lead to “averaged” points that deviate from the intrinsic data manifold, especially in cases where the intrinsic manifold is very curved (Extended Data Figure 1a). As cellular state spaces can be heavily non-linear [13, 16, 18], we require an alternative method of diffusion condensation that ensures that the condensed points remain on the manifold. A straightforward method for achieving this might be diffusion map coordinates. However, the computation of diffusion map coordinates requires eigendecomposition of a diffusion operator, which is known to be slow ( $O(n^3)$  complexity). In the current manuscript rather than using the original features, we use the potential representation of data points used in PHATE (see Equation 1) as the as initial features. This effectively re-represents points by features that consist of the log of diffusion probabilities to all other features. In PHATE, this representation is used as the step before dimensionality reduction and low-dimensional data embedding. However, we use these diffusion potential coordinates here as a high-dimensional representation of the data on which the condensation operates, offering a “straightened” and globally coherent intrinsic manifold space upon which to operate the diffusion condensation process. This way when data points are condensed, they are condensed in terms of their diffusion probabilities.

### 5.1.4 Scalable coarse-graining with fast diffusion condensation

In its original form, the condensation process proceeds as follows. Let  $\mathbf{X}^{(0)} = \mathbf{X}$  be the initial data matrix with diffusion operator  $\mathbf{P}_0 := \mathbf{P}$  constructed from its rows (as data points), and let  $\mathbf{X}^{(1)} = \mathbf{P}_0 \mathbf{X}^{(0)}$ . This gives the first iteration of the process, where the application of the diffusion smoothing intrinsically denoises and reduces local variability in  $\mathbf{X}^{(1)}$  compared to  $\mathbf{X}^{(0)}$ . Then, the process is repeated to further reduce local data variability by computing the diffusion operator  $\mathbf{P}_1$  over rows of  $\mathbf{X}^{(1)}$ , yielding  $\mathbf{X}^{(2)} = \mathbf{P}_1 \mathbf{X}^{(1)}$ . In general, this process is repeated iteratively, resulting in a time-inhomogeneous Markov process

$$\mathbf{X}^{(t+1)} = \mathbf{P}_t \mathbf{X}^{(t)} = \mathbf{P}_t \mathbf{P}_{t-1} \cdots \mathbf{P}_1 \mathbf{P}_0 \mathbf{X}, \quad (3)$$

whose  $t$ -step transition probabilities are given by the entries of a time-varying row-stochastic operator  $\mathbf{P}^{(t)} = \mathbf{P}_t \cdots \mathbf{P}_0$  (note that  $\mathbf{P}^{(t)}$  does not represent a matrix power, but rather the diffusion condensation operator at iteration  $t$ ). As mentioned previously, due to the low-pass nature of each diffusion operator  $\mathbf{P}_t$ , this Markov process adaptively removes local (high frequency) variations in input coordinate functions. The effect on the data points  $\mathbf{X}$  is to draw them towards local barycenters, which are defined by the inhomogeneous diffusion process.

In order to allow Multiscale PHATE to enable scalable exploration of large data sets, such as high dimensional biological data, we propose speeding up of the initial condensation iteration in the following ways:

1. Speed-up of the initial iteration using graph partitioning.
2. Fast computation of the diffusion potential via landmarking.
3. Merging of data points to increase computational efficiency over iterations

The complexity of computing a diffusion operator on  $n$  points is  $n^2$ . In order to reduce  $n$  for initial condensation iterations, we run hierarchical kmeans on the PCA space of the data with a high  $K$  (by default 100) to obtain a coarse graining of the data in feature space. In each iteration of the kmeans approach we partition the data into  $k$  more clusters. In subsequent iterations we compute another  $k$  clusters from each of these clusters. This process continues until we have a large number of clusters from which to compute the diffusion operator (by default 25,000). We then compute a landmarked diffusion potential (as done in [13] and explained below) on this reduced space, by convention the centroid of each of these clusters, before starting the coarse graining process.

Creation of the diffusion operator requires the computation of all pairwise distances between points, before conversion of those distances to affinities. Instead of using spectral clustering on the full dataset, we can come up with cluster centroids that are treated as "landmarks". Afterwards, distances  $\mathbf{D}_{pl}$  and  $\mathbf{A}_{pl}$  are computed between points and landmarks, i.e., they are  $n \times k$  matrices where  $n$  is the number of points and  $k$  is the number of landmarks. In addition, distances  $D_l$  and affinities  $\mathbf{A}_l$  are computed between landmarks, resulting in  $k \times k$  matrices. In order to compute the diffusion operator  $\mathbf{P}^t$ , we row normalize  $\mathbf{A}_{pl}$  and  $\mathbf{A}_l$  to obtain  $\mathbf{P}_{pl}$ ,  $\mathbf{P}_l$  and compute  $\mathbf{P}^t = \mathbf{P}_{pl} \mathbf{P}_l^t \mathbf{P}_{lp}$ , thus decomposing  $t$ -step path probabilities between two points as the probability of going to a landmark and then back to the point. We have shown in [13] that this leads to high quality approximations of the diffusion operator which lead to near-identical visualizations with PHATE. In addition, we examined in [84] that this leads to low error approximations of diffusion operators in general. We use this fast approach to compute a low error diffusion potential system for our coarse graining process.

In order to increase computational efficiency over successive iterations of condensation, we merge points that fall within a threshold distance into a single point. When two or more points collapse into the same barycenter (closer than a threshold  $\zeta$ ), we merge them into a cluster since they would then have approximately the same coordinates. After this merging operation, we effectively treat the cluster as a single point. Intuitively, this merging process creates a single connected component from two different components in our calculation of data topology. This has the effect of density subsampling the data iteratively, and allowing for subsequent iterations to proceed faster. Therefore the number of points steadily decreases, allowing the algorithm to speed up in successive iterations.

As we iterate this process over and over again, the condensation process slowly coarse grains the data to reveal structure at all levels of granularity while avoiding the typical tendency of traditional hierarchical clustering approaches to force (e.g., greedy) cluster merges at every scale.

We show that the resultant method is orders of magnitude faster than competing methods: DM, t-SNE, UMAP, Monocle2, and PHATE (Extended Data Figure 1c).

### 5.1.5 Selection of visualization layers via Gradient Analysis

As our iterative coarse graining approach creates hundreds of layers for downstream analysis, selecting salient level of granularity for visualization is a critical task. We envision users to start with a coarse-grained level that offers a high-level summarization of the data before "zooming-in" to obtain finer detail on populations of interest. To offer this type of interactivity, we provide a method for selecting specific scales for visualization.

We reason that the salient levels of the representation for visualization must be stable levels which emerge during condensation, i.e., levels whose structure persists for several iterations. To



find such levels we examine the gradient of points across successive condensation iterations and determine where the overall shift in data density from one iteration to the next is locally minimal (Figure 1b). In order to identify these metastable states, we identify changes in manifold density for every pair of successive condensation iterations  $\nabla^{(t,t+1)} = \mathbf{X}^{(t)} - \mathbf{X}^{(t+1)}$ . In order to identify total shifts in density we compute the matrix sum of  $\nabla^{(t,t+1)}$  by  $G = \sum_n^N \sum_n^N \nabla^{(t,t+1)}$ . Picking scales for visualization and downstream analysis arises from identifying local minima in  $G$  (Figure 1b). Visualization of a granularity is achieved by the stress-minimizing optimization procedure provided by multidimensional scaling (MDS) on the condensed diffusion potential as done in [13]. Finally, to obtain more refined and detailed visualizations, we allow users to select data subsets to view at finer granularities identified by gradient analysis (Figure 1c).

### 5.1.6 Distinction and comparison between the diffusion condensation process and hierarchical clustering

One use of diffusion condensation can be to provide a hierarchy of clusters determined by merged points. However, it should be noted that the condensation process here is significantly different from typical hierarchical clustering, and instead provides a richer coarse graining of data geometry. Indeed, hierarchical clustering algorithms generally belong to two families: divisive algorithms and agglomerative ones.

Divisive approaches (e.g., bisecting  $k$ -means [85] or MST-based clustering [86]) work in a top-down fashion, each time optimizing a partition of the data into clusters (e.g., using partitional methods like  $k$ -means), and then recursively partitioning further each cluster. The difference between these and the gradual aggregation approach of the condensation process is clear.

Agglomerative methods, on the other hand, work in a bottom-up fashion by first merging points into clusters, and then recursively merging increasingly larger clusters. While intuitively more related to the gradual merges in diffusion condensation, there is a fundamental difference between the coarse graining operation applied here and the (typically greedy) agglomeration in such methods. Indeed, most agglomerate clustering methods only operate on determining an iterative or recursive sequence of merges, without considering any intermediate information or structure in the data.

The condensation process utilized here, on the other hand, is derived from a continuous process that gradually eliminates local variability in the data. At its core, it relies on a time-inhomogeneous Markov chain that gradually constructs a diffusion geometry that reveals global and local structures in the data at increasingly coarse scales. The elimination of local variability in this process allows points to naturally come together, thus producing natural data clusters from data regions that collapse to the same point, without the need for partitioning or greedy agglomeration. However, this is a pattern that emerges from the coarse graining process, rather than directly or explicitly guiding it. The constructed multiresolution data geometry also reveals other information, beyond clustering, which makes it amenable for visualization and other downstream tasks. For instance, condensation homology produces persistent features that are meaningful, and levels of meta-stability can be analyzed, as we do for the selection of meta-stable resolutions (e.g., for visualization) explained below.

To demonstrate the difference between diffusion condensation and agglomerative clustering, we use the Louvain method [20] as a representative example, due to its popularity in single cell data analysis. This method greedily selects clusters to merge together by their impact on modularity (i.e., whether and how much they improve it). The forced merges, while ensuring a hierarchy of data agglomerations, do not provide reliable coarse grained representations for revealing varied data resolutions. As we showed in Extended Data Figure 3, they miss vital levels of resolution. Meanwhile, diffusion condensation allows for a systematic exploration of granularity and is better at capturing

levels where biological differences may exist (Extended Data Figure 3e).

In order to quantitatively compare the accuracy of Multiscale PHATE clusters with hierarchical clustering approaches, we compared cluster labels generated from a range of clustering strategies to ground truth labels using Adjusted Rand index (ARI). We first generated a hierarchical stochastic block model with different clusters at multiple granularities (Extended Data Figure 3a). We then used Multiscale PHATE, Louvain [20], Leiden [21] and single linkage hierarchical clustering [22] to identify groupings across multiple levels of granularity. For each level of ground truth clusters, we computed ARI against cluster labels from each algorithm across all granularities, storing the highest ARI for each method. For the flow cytometry data, we used gated populations from 3 samples in our myeloid-centric flow cytometry panel as ground truth labels across coarse and fine grain cluster labels. For instance, at coarse grain monocytes would be identified as one population, however at fine grain monocytes would be a part of three distinct populations. ARI was computed similarly for this dataset, ground truth labels were compared to all granularities of clusters from each algorithm, with the top score stored for each approach. Networkx [87] was used to produce Louvain clusters, leidenalg was used to produce Leiden clusters and agglomerative clusters were produced using sklearn [88].

### 5.1.7 Comparison of multigranular visualizations with DeMAP

DeMAP is a metric for assessing visualization quality in terms of its ability to capture the manifold geometry of noisy data [13]. DeMAP computes correlation between geodesic distances on ground truth noiseless data manifolds to Euclidean distances on embedding created from noisy data. High DeMAP scores indicate visualization that accurately represents geodesic manifold distances in an embedding.

In order to show that Multiscale PHATE created improved multigranular visualizations when compared to other approaches, we performed two ablation studies. First, the *splatter* software was used to simulate ground truth and noisy single cell data of either group (cluster) or path (trajectory) geometries [19]. In the first ablation study, differing approaches to build a multiscale abstraction of the noisy synthetic data were computed: louvain and computational homology. These abstractions were visualized with a range of visualization strategies. The resultant embeddings were compared with multiscale PHATE using DeMAP. In the second ablation study, condensation topology on the noisy synthetic data was computed and a multiscale PHATE embedding was created after identifying the optimal resolution via gradient analysis. In order to create multiscale visualizations with other dimensionality reduction strategies, we applied a range of other visualization approaches to this condensed granularity of noisy data. Finally, all embeddings were compared using DeMAP. These studies were repeated across a range of noise types, biological variation and drop out, and a range of noise levels. For robustness, this process was run across 10 different splatter datasets with group geometry and 10 different splatter datasets with path geometry for each comparison. Besides Multiscale PHATE, the DeMAP package was used to build all other visualizations [13].

### 5.1.8 Construction of patient manifold through multiresolution cluster evaluation:

Previously, PhEMD built a manifold of samples measured via single cell technologies by binning cells associated with each sample into histograms and computing Earth Mover Distances or optimal transport between histograms [89]. This computation, however, is done at a single scale and requires determining ground distance between histogram bins. Truthfully, cells can occupy a diverse set of hierarchical labels which a single resolution does not capture. Recently, these optimal transport based techniques have been implemented on hierarchical trees allowing for multiresolution comparison of



sample density variations [23]. Furthermore, in [90] it was shown that multiscale smoothing of data distribution can be used to approximate the Earth-mover distances between them, which are closely related to optimal transport and essentially measures the energy required to transform one data distribution to another. Intuitively, this can be understood as computing how difficult it would be to change the underlying cellular state of one patient to that of another patient.

Inspired by these techniques, we have developed a multiresolution evaluation system for determining sample level density variations to build a manifold of patients based on differences in their underlying cellular states. Here, instead, we combine multiple histograms at different scales for each patient to compute a distance between their underlying cellular states. Our approach of replacing ground distance with multiscale construction is based on the work of [90] and [23], which show that with appropriate weights, the combination of such smoothed data distributions can be used to efficiently compute or approximate Earth Mover distances. We note that our results show empirically that even without careful tuning of such weights, the resulting patient to patient distance, and the constructed manifold, accurately recapitulate the clinical states.

Practically, we create a manifold of samples by simultaneously evaluating multiple levels of the diffusion condensation topology. At each level  $\ell \in \{1, 2, \dots, L\}$ , a number of  $N_\ell$  clusters are identified. We count the number of cells,  $n_{\ell,j,k}$ , of the  $k$ -th patient that belong to each cluster  $C_{\ell,j}$  for every  $j \in \{1, 2, \dots, N_\ell\}$  and calculate the normalized percentage as  $r_{\ell,j,k} = \frac{n_{\ell,j,k}}{\sum_j n_{\ell,j,k}}$ . We calculate the proportions for all patients at a series of selected levels of the topology and concatenate these to create a rich multiscale vector of features for each patient. These multiscale feature vectors are then used to create an embedding with PHATE [13] and to de-noise patient specific signals using MAGIC [18] using Euclidean distance between samples.

### 5.1.9 Use of MELD with Multiscale PHATE

MELD is a method proposed in [15] that takes a discrete signal defined on a data graph and computes a continuous likelihood score of the signal value by using a sophisticated form of neighborhood averaging by using a heat kernel at each point (Figure 1c). In order to apply MELD to this dataset, we combined the flow cytometry data coming from all patients, and used a binary outcome score that we call *mortality*, which uses a discrete 0-value for a positive outcome (the patient was discharged), or a 1-value for a negative outcome (patient died or was sent to hospice). The outcome of the patient is used as the discrete condition for all cells from that patient. Thus in our combined flow-cytometry dataset, every cell from positive outcome patients get a raw experimental signal value of 0. Using MELD, we estimate the likelihood of each outcome over the cellular manifold using a heat-diffusion kernel applied to the data graph to obtain mortality likelihood score. Values of the mortality likelihood score range from 0 to 1 and constitute a probability likelihood estimate of the condition over the manifold. This allows us to identify areas of the cellular manifold that are likely to be enriched in those with positive or negative outcomes.

Since Multiscale PHATE identifies clusters of cells across all levels of granularity, we can sweep across resolutions to identify levels which isolate high and low mortality likelihood score regions. In fact, when comparing our multigranular clusters with other clustering techniques across a range of granularities, we show that multiscale PHATE is better able to isolate high and low mortality likelihood score regions in one of our flow cytometry panels (Extended Data Figure 3e). By looking at these informative resolutions, we can identify populations of cells that are pertinent to patient outcomes. When identifying these subpopulations in conjunction with cell type defining markers, we show that we can identify cell types and functional subtypes that are differentially enriched across patient outcomes and may drive disease pathogenesis. The full Multiscale PHATE and MELD integrated pipeline can be seen in Extended Data Figure 1b.

#### 5.1.10 DREMI Associations with mortality likelihood score

DREMI [16], is an information-theoretic metric that quantifies associations or strength of a relationship between two variables. Like most discrete estimates of mutual information, DREMI starts by binning continuous data into equal-sized partitions,  $X = \{X_1, X_2, \dots, X_n\}$ , and  $Y = \{Y_1, Y_2, \dots, Y_n\}$  in both variable dimensions but instead of measuring the mutual information as  $I(X, Y) = H(Y) - \sum_i H(Y|X_i)$  the difference between the entropy of  $Y$  and the conditional entropy of  $X|Y$ , DREMI "resamples" or equalizes the number of samples in each bin using an extra level of conditioning. Thus DREMI computes  $DREMI(X, Y) = I(X, Y|X) = H(Y|X) - \sum_i H(Y|X_i)$ . The rationale for this is that normal mutual information is dominated by the density peaks of the  $X$  variable, and does not reveal the full strength of the relationship given imbalanced sampling which is common in biomedical data.

When combining our DREMI analysis with previously computed mortality likelihood score, we can identify functional marker trends which are correlated with mortality. As cells of the same type can occupy a range of functional states that can be enriched in disease, a given subtype may not be associated with mortality but a functional substate could be. By computing DREMI associations between mortality likelihood score and cellular functional state markers, we can identify markers, and by extension activation states, that are associated with outcome.

#### 5.1.11 Patient manifold analysis from Multiscale PHATE Features

In order to identify the differences between individual patient samples, we used Multiscale PHATE to construct a manifold of patients as described above. Similar to mortality likelihood score computed by MELD in our flow cytometry analysis, we computed a similar mortality likelihood score for our patient manifold by identifying if each patient sample originated from a patient that had a positive outcome or a negative outcome. In order to identify patient sample features correlated with mortality likelihood score, we compiled a set of clinical, demographic and Multiscale PHATE identified cell type proportion features for each patient sample. Using the geometry of the patient manifold, we de-noised our patient sample features using MAGIC [18] before running association analysis between features using DREMI [16].

#### 5.1.12 Mortality Prediction using Random Forest Classifier

In addition to being useful for visualizing, clustering and identifying condition specific enrichment of cell types, we wanted to see if the populations we identified across granularities were predictive of patient outcome. In order to predict patient outcomes from just a single patient sample, we trained a random forest classifier on populations we identified in our myeloid focused flow cytometry panel. Similar to our patient manifold analysis, we derived multiscale patient features by identifying the proportion of each patient's cells that were labeled with a particular cell type. After partitioning our dataset of 210 patient samples into 5 sets, we performed 5-fold cross-validation where we iteratively shuffled training sets (4 of 5) and test sets (1 of 5). Across all runs, we achieved an accuracy of 83.5% across both conditions, with 86.5% classification accuracy for patients that survive and 78.8% classification accuracy for patients that died from infection. In order to identify cellular types that were particularly informative of mortality outcome, we computed and compared the feature importance of random forests. This analysis revealed that monocytes, CD16<sup>+</sup> neutrophils, dendritic cells and T cells had the highest importance and were most predictive of mortality outcome. To determine whether our Multiscale PHATE derived cellular populations were more informative than current gold standard cell typing strategies, we also trained a random forest classifier on cell

populations identified via conventional flow cytometry gating analysis. This analysis was only able to accurately predict patient outcomes in 73.8% of cases.

### 5.1.13 Software availability

The Multiscale PHATE package, as implemented in python, is available for download with a guided tutorial on the Krishnaswamy Lab Github page: [https://github.com/KrishnaswamyLab/Multiscale\\_PHATE](https://github.com/KrishnaswamyLab/Multiscale_PHATE).

### 5.1.14 Pre-processing of patient flow cytometry data

Flow cytometry was performed on PBMC from each patients (methods explained in detail below). The resulting .FCS files were pre-processed by applying compensation based on the respective single-color compensation controls, 2) selecting only leukocytes and singlets based on FSC and SSC, and 3) selecting only live cells based on a viability dye. MFI values for each fluorophore on a per-cell basis were then extracted for downstream analysis. In order to extract T cells for the cytokine focused T cell panel, cells with CD3 staining greater than 425 were extracted. For the T cell surface marker panel, cells with a CD3 staining greater than 500 were extracted. For the B cell focused panel, cells with a CD19 staining greater than 400 were extracted and cells expressing less than a total of 2700 cumulative staining across all markers were removed. No extraction of cells was done for the myeloid focused panel, however cells with cumulative staining across all markers less than 2700 across were removed. All datasets were then independently normalized to 1000 staining counts per cell before square root normalization.

## 5.2 Biological and Medical Methods

In the following sections we provide details on how patient biological data and clinical information was acquired and processed.

### 5.2.1 Ethics statement

This study was approved by Yale Human Research Protection Program Institutional Review Boards (FWA00002571, protocol ID 2000027690). Informed consent was obtained from all enrolled patients and healthcare workers.

### 5.2.2 Patients

Patient enrollment, sample acquisition, processing, and downstream analysis by flow cytometry were performed as in [14]. One-hundred and sixty-eight patients admitted to YNHH with SARS-CoV2 between 18 March 2020 and 27 May 2020 were recruited to the Yale IMPACT study (Implementing Medical and Public Health Action Against Coronavirus CT) after testing positive for SARS-CoV2 by qRT-PCR and included in this study. No statistical methods were used to predetermine sample size. Paired whole blood for flow cytometry analysis was collected simultaneously in sodium heparin-coated vacutainers and kept on gentle agitation until processing. All blood was processed on the day of collection. Patients were scored for COVID-19 disease severity through review of electronic medical records (EMR) at each longitudinal time point. For all patients, days from symptom onset were estimated as follows: (1) highest priority was given to explicit onset dates provided by patients; (2) next highest priority was given to the earliest reported symptom by a patient; and (3) in the absence of direct information regarding symptom onset, we estimated a date through

manual assessment of the electronic medical record (EMRs) by an independent clinician. The clinical data were collected using EPIC EHR and REDCap 9.3.6 software. At the time of sample acquisition and processing, investigators were unaware of the patients' conditions. Blood acquisition was performed and recorded by a separate team. Information about patients' conditions was not available until after processing and analysis of raw data by flow cytometry and ELISA. A clinical team, separate from the experimental team, performed chart reviews to determine relevant statistics. Flow cytometry analyses were performed blinded. Patients' clinical information and clinical score coding were revealed only after data collection.

### 5.2.3 Isolation of PBMCs

PBMCs were isolated from heparinized whole blood using Histopaque (Sigma-Aldrich, #10771-500ML) density gradient centrifugation in a biosafety level 2+ facility. After isolation of undiluted serum, blood was diluted 1:1 in room temperature PBS, layered over Histopaque in a SepMate tube (StemCell Technologies; #85460) and centrifuged for 10 min at 1,200g. The PBMC layer was isolated according to the manufacturer's instructions. Cells were washed twice with PBS before counting. Pelleted cells were briefly treated with ACK lysis buffer for 2 min and then counted. Percentage viability was estimated using standard Trypan blue staining and an automated cell counter (Thermo-Fisher, #AMQAX1000).

### 5.2.4 Flow cytometry

In brief, freshly isolated PBMCs were plated at  $1-2 \times 10^6$  cells per well in a 96-well U-bottom plate. Cells were resuspended in Live/Dead Fixable Aqua (ThermoFisher) for 20 min at 4 °C. Following a wash, cells were blocked with Human TruStain FcX (BioLegend) for 10 min at RT. Cocktails of desired staining antibodies were added directly to this mixture for 30 min at RT. For secondary stains, cells were first washed and supernatant aspirated; then to each cell pellet a cocktail of secondary markers was added for 30 min at 4 °C. Prior to analysis, cells were washed and resuspended in 100µL of 4% PFA for 30 min at 4 °C. For intracellular cytokine staining following stimulation, cells were resuspended in 200µL cRPMI (RPMI-1640 supplemented with 10% FBS, 2 mM l-glutamine, 100 U/ml penicillin, and 100 ug/ml streptomycin, 1 mM sodium pyruvate, and 50µM 2-mercaptoethanol) and stored at 4 °C overnight. Subsequently, these cells were washed and stimulated with 1× Cell Stimulation Cocktail (eBioscience) in 200 µL cRPMI for 1 h at 37 °C. 50µL of 5x Stimulation Cocktail (plus protein transport inhibitor) (eBioscience) was added for an additional 4 h of incubation at 37 °C. Following stimulation, cells were washed and resuspended in 100 µL of 4% PFA for 30 min at 4 °C. To quantify intracellular cytokines, these samples were permeabilized with 1× permeabilization buffer from the FOXP3/Transcription Factor Staining Buffer Set (eBioscience) for 10 min at 4 °C. All subsequent staining cocktails were made in this buffer. Permeabilized cells were then washed and resuspended in a cocktail containing Human TruStain FcX (BioLegend) for 10 min at 4 °C. Finally, intracellular staining cocktails were added directly to each sample for 1 h at 4 °C. Following this incubation, cells were washed and prepared for analysis on an Attune NXT (ThermoFisher). Data were analysed using FlowJo software version 10.6 software (Tree Star).

### 5.2.5 Acquisition of Clinical Data for Flow Cytometry analysis and Patient Manifest

Longitudinal patient data was extracted from the electronic medical record (Epic, Verona, WI) for only the hospitalized patients included in the repository. Time-varying data, specifically vital signs as well as laboratory studies, were extracted specifically 24 hours before and after the collection of blood specimens for flow cytometry as described above. This ensured that the measurements correlated

with the patient state at the time of flow cytometry measurements. Laboratory values reflecting clinical evaluation of general inflammatory states (white blood cell count, high sensitivity c-reactive protein) were extracted. The values for the laboratory measurements were then consolidated by taking the most abnormal value (e.g. highest ferritin) in the 72 hour period and overlaid onto the patient manifolds.

### 5.2.6 Acquisition of Clinical Data for Clinical Manifold

For patients who did not undergo flow cytometry analysis, the time varying clinical, laboratory, and treatment data was extracted for the first 24 hours from admission with consolidation by the most abnormal value as described before. Otherwise, the consolidated data temporally correlating to flow cytometry measurements were extracted as described above.

## 6 Acknowledgements

Yale IMPACT Research Team

Abeer Obaid<sup>16</sup>, Adam Moore<sup>21</sup>, Alice Lu-Culligan<sup>3</sup>, Allison Nelson<sup>16</sup>, Anderson Brito<sup>10</sup>, Angela Nunez<sup>16</sup>, Anjelica Martin<sup>3</sup>, Anne L Wyllie<sup>9</sup>, Annie Watkins<sup>10</sup>, Annsea Park<sup>3</sup>, Arvind Venkataraman<sup>3</sup>, Bertie Geng<sup>16</sup>, Chaney Kalinich<sup>10</sup>, Chantal BF Vogels<sup>9</sup>, Christina Harden<sup>10</sup>, Codruta Todeasa<sup>16</sup>, Cole Jensen<sup>10</sup>, Daniel Kim<sup>3</sup>, David McDonald<sup>16</sup>, Denise Shepard<sup>16</sup>, Edward Courchaine<sup>17</sup>, Elizabeth B. White<sup>10</sup>, Eric Song<sup>3</sup>, Erin Silva<sup>16</sup>, Eriko Kudo<sup>3</sup>, Giuseppe DeIuliis<sup>12</sup>, Haowei Wang<sup>10</sup>, Harold Rahming<sup>16</sup>, Hong-Jai Park<sup>16</sup>, Irene Matos<sup>16</sup>, Isabel M Ott<sup>9</sup>, Jessica Nouws<sup>16</sup>, Jordan Valdez<sup>16</sup>, Joseph Fauver<sup>10</sup>, Joseph Lim<sup>18</sup>, Kadi-Ann Rose<sup>16</sup>, Kelly Anastasio<sup>19</sup>, Kristina Brower<sup>10</sup>, Laura Glick<sup>16</sup>, Lokesh Sharma<sup>16</sup>, Lorenzo Sewanan<sup>16</sup>, Lynda Knaggs<sup>16</sup>, Maksym Minasyan<sup>16</sup>, Maria Batsu<sup>16</sup>, Maria Tokuyama<sup>3</sup>, M. Cate Muenker<sup>16</sup>, Mary Petrone<sup>10</sup>, Maxine Kuang<sup>10</sup>, Maura Nakahata<sup>16</sup>, Melissa Campbell<sup>15</sup>, Melissa Linehan<sup>3</sup>, Michael H. Askenase<sup>20</sup>, Michael Simonov<sup>16</sup>, Mikhail Smolgovsky<sup>16</sup>, Nathan D. Grubaugh<sup>25</sup>, Nicole Sonnert<sup>3</sup>, Nida Naushad<sup>16</sup>, Pavithra Vijayakumar<sup>16</sup>, Peiwen Lu<sup>3</sup>, Rebecca Earnest<sup>9</sup>, Rick Martinello<sup>1</sup>, Roy Herbst<sup>16,23,24</sup>, Rupak Datta<sup>1</sup>, Ryan Handoko<sup>16</sup>, Santos Bermejo<sup>16</sup>, Sarah Lapidus<sup>9</sup>, Sarah Prophet<sup>16</sup>, Sean Bickerton<sup>17</sup>, Sofia Velazquez<sup>20</sup>, Subhasis Mohanty<sup>10</sup>, Tara Alpert<sup>1</sup>, Tyler Rice<sup>3</sup>, Wade Schulz<sup>22</sup>, William Khoury-Hanold<sup>3</sup>, Xiaohua Peng<sup>16</sup>, Yexin Yang<sup>3</sup>, Yiyun Cao<sup>3</sup> & Yvette Strong<sup>16</sup>

<sup>16</sup>Yale University School of Medicine, New Haven, CT, USA. <sup>17</sup>Department of Biochemistry and Molecular Biology, Yale University School of Medicine, New Haven, CT, USA. <sup>18</sup>Yale Viral Hepatitis Program, Yale University School of Medicine, New Haven, CT, USA. <sup>19</sup>Yale Center for Clinical Investigation, Yale University School of Medicine, New Haven, CT, USA. <sup>20</sup>Department of Neurology, Yale University School of Medicine, New Haven, CT, USA. <sup>21</sup>Yale University School of Public Health, New Haven, CT, USA. <sup>22</sup>Center of Biomedical Data Science, Yale University, New Haven, CT, USA. <sup>23</sup>Yale Cancer Center, Yale New Haven Hospital, CT, USA. <sup>24</sup>Smilow Cancer Hospital, Yale New Haven Hospital, New Haven, CT, USA. <sup>25</sup>Department of Epidemiology of Microbial Diseases, Yale University School of Public Health, New Haven, CT, USA.

## 7 Author Contributions

Conception: M.K., J.H., P.W., J.G., S.K. A.I., G.W., J.H., S.F., C.S.D., A.I.K., P.W. ; Design of Work: M.K., J.H., J.G., D.S., A.T. S.K. A.I., G.W., J.H. ; Acquisition of Data: P.W., J.G., C.L., J.K., B.I., M.S., T.M., J.E.O., J.S., T.T., C.D.O., A.C., J.F. ; Analysis of Data: M.K., J.H., P.W., J.G., D.B.B., A.T., S.G., A.G., B.R. ; Interpretation of Data: M.K., J.H., P.W., J.G., S.K. A.I.,

G.W., J.H., S.F., C.S.D., A.I.K., P.W. ; Creation of New Software: M.K., J.H. ; Writing - Drafting: M.K., J.H., P.W. ; Writing - Revision: S.K. A.I., G.W., J.H., S.F., C.S.D., A.I.K., P.W. ;

## 8 Competing Interests

Dr. Krishnaswamy is on the scientific advisory board of KovaDx and AI Therapeutics.

Dr. Iwasaki a member of the SAB for InProTher.

Dr. Iwasaki is a co-founder of RIGImmune.

Dr. Wilson is founder of Efference.

Dr. Ko is a member of the expert panel of the Reckit Global Hygiene Institute.

## References

- [1] Bandura, D. R. *et al.* Mass cytometry: technique for real time single cell multitarget immunoassay based on inductively coupled plasma time-of-flight mass spectrometry. *Analytical chemistry* **81**, 6813–6822 (2009).
- [2] Spitzer, M. H. & Nolan, G. P. Mass cytometry: single cells, many features. *Cell* **165**, 780–791 (2016).
- [3] Brummelman, J. *et al.* Development, application and computational analysis of high-dimensional fluorescent antibody panels for single-cell flow cytometry. *Nature protocols* **1** (2019).
- [4] Klein, A. M. *et al.* Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
- [5] Macosko, E. Z. *et al.* Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
- [6] Buenrostro, J. D. *et al.* Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
- [7] van der Maaten, L. & Hinton, G. Visualizing high-dimensional data using t-sne. *Journal of Machine Learning Research* **9**, 2579–2605 (2008).
- [8] Becht, E. *et al.* Dimensionality reduction for visualizing single-cell data using UMAP. *Nature Biotechnology* **37**, 38 (2019).
- [9] Pearson, K. LIII. on lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science* **2**, 559–572 (1901). URL <https://doi.org/10.1080/14786440109462720>.
- [10] Lee, J. S. *et al.* Immunophenotyping of covid-19 and influenza highlights the role of type i interferons in development of severe covid-19. *Science Immunology* **5** (2020). URL <https://immunology.sciencemag.org/content/5/49/eabd1554>. <https://immunology.sciencemag.org/content/5/49/eabd1554.full.pdf>.
- [11] Liao, M. *et al.* Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nature Medicine* **26**, 842–844 (2020). URL <https://doi.org/10.1038/s41591-020-0901-9>.

- [12] Brugnone, N. *et al.* Coarse graining of data via inhomogeneous diffusion condensation. In *2019 IEEE International Conference on Big Data (Big Data)*, 2624–2633 (IEEE, 2019).
- [13] Moon, K. R. *et al.* Visualizing structure and transitions in high-dimensional biological data. *Nature Biotechnology* **37**, 1482–1492 (2019).
- [14] Lucas, C. *et al.* Longitudinal analyses reveal immunological misfiring in severe COVID-19. *Nature* **584**, 463–469 (2020). URL <https://doi.org/10.1038/s41586-020-2588-y>.
- [15] Burkhardt, D. B. *et al.* Quantifying the effect of experimental perturbations in single-cell rna-sequencing data using graph signal processing. *bioRxiv* (2020). URL <https://www.biorxiv.org/content/early/2020/08/01/532846>. <https://www.biorxiv.org/content/early/2020/08/01/532846.full.pdf>.
- [16] Krishnaswamy, S. *et al.* Conditional density-based analysis of t cell signaling in single-cell data. *Science* **346**, 1250689–1250689 (2014).
- [17] Carlsson, G. Topological pattern recognition for point cloud data. *Acta Numerica* **23**, 289–368 (2014).
- [18] van Dijk, D. *et al.* Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716 – 729.e27 (2018).
- [19] Zappia, L., Phipson, B. & Oshlack, A. Splatter: simulation of single-cell RNA sequencing data. *Genome Biology* **18** (2017). URL <https://doi.org/10.1186/s13059-017-1305-0>.
- [20] Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *Journal of Statistical Mechanics: Theory and Experiment* **2008**, P10008 (2008).
- [21] Traag, V. A., Waltman, L. & van Eck, N. J. From louvain to leiden: guaranteeing well-connected communities. *Scientific Reports* **9** (2019). URL <https://doi.org/10.1038/s41598-019-41695-z>.
- [22] Sibson, R. SLINK: An optimally efficient algorithm for the single-link cluster method. *The Computer Journal* **16**, 30–34 (1973). URL <https://doi.org/10.1093/comjnl/16.1.30>.
- [23] Le, T., Yamada, M., Fukumizu, K. & Cuturi, M. Tree-Sliced Variants of Wasserstein Distances. *NeurIPS* (2019). [1902.00342](https://arxiv.org/abs/1902.00342).
- [24] Marshall, J. C. *et al.* A minimal common outcome measure set for COVID-19 clinical research. *The Lancet Infectious Diseases* **20**, e192–e197 (2020). URL [https://doi.org/10.1016/s1473-3099\(20\)30483-7](https://doi.org/10.1016/s1473-3099(20)30483-7).
- [25] Garley, M. & Jabłońska, E. Heterogeneity among neutrophils. *Archivum Immunologiae et Therapiae Experimentalis* **66**, 21–30 (2017). URL <https://doi.org/10.1007/s00005-017-0476-4>.
- [26] Wang, Y. *et al.* ADAM17 cleaves CD16b (fcγRIIIb) in human neutrophils. *Biochimica et Biophysica Acta (BBA) - Molecular Cell Research* **1833**, 680–685 (2013). URL <https://doi.org/10.1016/j.bbamcr.2012.11.027>.
- [27] Pillay, J. *et al.* A subset of neutrophils in human systemic inflammation inhibits t cell responses through mac-1. *Journal of Clinical Investigation* **122**, 327–336 (2012). URL <https://doi.org/10.1172/jci57990>.



- [28] Fortunati, E., Kazemier, K. M., Grutters, J. C., Koenderman, L. & van J. M. M. Van den Bosch. Human neutrophils switch to an activated phenotype after homing to the lung irrespective of inflammatory disease. *Clinical & Experimental Immunology* **155**, 559–566 (2009). URL <https://doi.org/10.1111/j.1365-2249.2008.03791.x>.
- [29] Padgett, L. E. *et al.* Interplay of monocytes and t lymphocytes in COVID-19 severity (2020). URL <https://doi.org/10.1101/2020.07.17.209304>. MedRxiv: 10.1101/2020.07.17.209304.
- [30] Sánchez-Cerrillo, I. *et al.* COVID-19 severity associates with pulmonary redistribution of CD1c+ DC and inflammatory transitional and nonclassical monocytes. *Journal of Clinical Investigation* (2020). URL <https://doi.org/10.1172/jci140335>.
- [31] Laing, A. G. *et al.* A dynamic COVID-19 immune signature includes associations with poor prognosis. *Nature Medicine* (2020). URL <https://doi.org/10.1038/s41591-020-1038-6>.
- [32] Winkler, M. S. *et al.* Human leucocyte antigen (HLA-DR) gene expression is reduced in sepsis and correlates with impaired TNF $\alpha$  response: A diagnostic tool for immunosuppression? *PLOS ONE* **12**, e0182427 (2017). URL <https://doi.org/10.1371/journal.pone.0182427>.
- [33] van der Poll, T., van de Veerdonk, F. L., Scicluna, B. P. & Netea, M. G. The immunopathology of sepsis and potential therapeutic targets. *Nature Reviews Immunology* **17**, 407–420 (2017). URL <https://doi.org/10.1038/nri.2017.36>.
- [34] Hynninen, M. *et al.* PREDICTIVE VALUE OF MONOCYTE HISTOCOMPATIBILITY LEUKOCYTE ANTIGEN-DR EXPRESSION AND PLASMA INTERLEUKIN-4 AND -10 LEVELS IN CRITICALLY ILL PATIENTS WITH SEPSIS. *Shock* **20**, 1–4 (2003). URL <https://doi.org/10.1097/01.shk.0000068322.08268.b4>.
- [35] Monneret, G. *et al.* The anti-inflammatory response dominates after septic shock: association of low monocyte HLA-DR expression and high interleukin-10 concentration. *Immunology Letters* **95**, 193–198 (2004). URL <https://doi.org/10.1016/j.imlet.2004.07.009>.
- [36] Lee, J. *et al.* The MHC class II antigen presentation pathway in human monocytes differs by subset and is regulated by cytokines. *PLOS ONE* **12**, e0183594 (2017). URL <https://doi.org/10.1371/journal.pone.0183594>.
- [37] Zhao, Y. *et al.* Longitudinal COVID-19 profiling associates IL-1ra and IL-10 with disease severity and RANTES with mild disease. *JCI Insight* **5** (2020). URL <https://doi.org/10.1172/jci.insight.139834>.
- [38] Mathew, D. *et al.* Deep immune profiling of COVID-19 patients reveals distinct immunotypes with therapeutic implications. *Science* **369**, eabc8511 (2020). URL <https://doi.org/10.1126/science.abc8511>.
- [39] Woodruff, M. *et al.* Dominant extrafollicular b cell responses in severe COVID-19 disease correlate with robust viral-specific antibody production but poor clinical outcomes (2020). URL <https://doi.org/10.1101/2020.04.29.20083717>. MedRxiv, DOI: 10.1101/2020.04.29.20083717.
- [40] Biasi, S. D. *et al.* Expansion of plasmablasts and loss of memory b cells in peripheral blood from COVID-19 patients with pneumonia. *European Journal of Immunology* **50**, 1283–1294 (2020). URL <https://doi.org/10.1002/eji.202048838>.

- [41] Kishimoto, T. Factors affecting b-cell growth and differentiation. *Annual Review of Immunology* **3**, 133–157 (1985). URL <https://doi.org/10.1146/annurev.iy.03.040185.001025>.
- [42] Robbiani, D. F. *et al.* Convergent antibody responses to SARS-CoV-2 in convalescent individuals. *Nature* **584**, 437–442 (2020). URL <https://doi.org/10.1038/s41586-020-2456-9>.
- [43] Kudva, A. *et al.* Influenza a inhibits th17-mediated host defense against bacterial pneumonia in mice. *The Journal of Immunology* **186**, 1666–1674 (2010). URL <https://doi.org/10.4049/jimmunol.1002194>.
- [44] Lee, Y. *et al.* Induction and molecular signature of pathogenic TH17 cells. *Nature Immunology* **13**, 991–999 (2012). URL <https://doi.org/10.1038/ni.2416>.
- [45] Skendros, P. *et al.* Complement and tissue factor-enriched neutrophil extracellular traps are key drivers in COVID-19 immunothrombosis. *Journal of Clinical Investigation* (2020). URL <https://doi.org/10.1172/jci141374>.
- [46] Zuo, Y. *et al.* Neutrophil extracellular traps in COVID-19. *JCI Insight* (2020). URL <https://doi.org/10.1172/jci.insight.138999>.
- [47] Middleton, E. A. *et al.* Neutrophil extracellular traps contribute to immunothrombosis in COVID-19 acute respiratory distress syndrome. *Blood* **136**, 1169–1179 (2020). URL <https://doi.org/10.1182/blood.2020007008>.
- [48] Sodhi, C. P. *et al.* A dynamic variation of pulmonary ACE2 is required to modulate neutrophilic inflammation in response to pseudomonas aeruginosa lung infection in mice. *The Journal of Immunology* **203**, 3000–3012 (2019). URL <https://doi.org/10.4049/jimmunol.1900579>.
- [49] Zhou, Z. *et al.* Heightened Innate Immune Responses in the Respiratory Tract of COVID-19 Patients. *Cell Host Microbe* **27**, 883–890 (2020).
- [50] Bost, P. *et al.* Host-Viral Infection Maps Reveal Signatures of Severe COVID-19 Patients. *Cell* **181**, 1475–1488 (2020).
- [51] Liang, S. C. *et al.* An IL-17f/a heterodimer protein is produced by mouse th17 cells and induces airway neutrophil recruitment. *The Journal of Immunology* **179**, 7791–7799 (2007). URL <https://doi.org/10.4049/jimmunol.179.11.7791>.
- [52] Jones, C. E. & Chan, K. Interleukin-17 stimulates the expression of interleukin-8, growth-related oncogene-  $\alpha$  , and granulocyte-colony-stimulating factor by human airway epithelial cells. *American Journal of Respiratory Cell and Molecular Biology* **26**, 748–753 (2002). URL <https://doi.org/10.1165/ajrcmb.26.6.4757>.
- [53] Liu, R. *et al.* IL-17 Promotes Neutrophil-Mediated Immunity by Activating Microvascular Pericytes and Not Endothelium. *J Immunol* **197**, 2400–2408 (2016).
- [54] Meierovics, A. I. & Cowley, S. C. MAIT cells promote inflammatory monocyte differentiation into dendritic cells during pulmonary intracellular infection. *J Exp Med* **213**, 2793–2809 (2016).
- [55] Channappanavar, R., Fett, C., Zhao, J., Meyerholz, D. K. & Perlman, S. Virus-specific memory CD8 t cells provide substantial protection from lethal severe acute respiratory syndrome coronavirus infection. *Journal of Virology* **88**, 11034–11044 (2014). URL <https://doi.org/10.1128/jvi.01505-14>.

- [56] Barber, D. L., Wherry, E. J. & Ahmed, R. Cutting edge: Rapid in vivo killing by memory CD8 t cells. *The Journal of Immunology* **171**, 27–31 (2003). URL <https://doi.org/10.4049/jimmunol.171.1.27>.
- [57] Kang, C. K. *et al.* Aberrant hyperactivation of cytotoxic t-cell as a potential determinant of COVID-19 severity. *International Journal of Infectious Diseases* **97**, 313–321 (2020). URL <https://doi.org/10.1016/j.ijid.2020.05.106>.
- [58] Moskophidis, D. & Kioussis, D. Contribution of virus-specific CD8+ cytotoxic t cells to virus clearance or pathologic manifestations of influenza virus infection in a t cell receptor transgenic mouse model. *Journal of Experimental Medicine* **188**, 223–232 (1998). URL <https://doi.org/10.1084/jem.188.2.223>.
- [59] Cannon, M. J., Openshaw, P. J. & Askonas, B. A. Cytotoxic t cells clear virus but augment lung pathology in mice infected with respiratory syncytial virus. *The Journal of Experimental Medicine* **168**, 1163–1168 (1988). URL <https://doi.org/10.1084/jem.168.3.1163>.
- [60] Bem, R. A. *et al.* Activation of the granzyme pathway in children with severe respiratory syncytial virus infection. *Pediatric Research* **63**, 650–655 (2008). URL <https://doi.org/10.1203/pdr.0b013e31816fdc32>.
- [61] Neidleman, J. *et al.* SARS-CoV-2-specific t cells exhibit phenotypic features of helper function, lack of terminal differentiation, and high proliferation potential. *Cell Reports Medicine* **1**, 100081 (2020). URL <https://doi.org/10.1016/j.xcrm.2020.100081>.
- [62] Hewagama, A., Patel, D., Yarlagadda, S., Strickland, F. M. & Richardson, B. C. Stronger inflammatory/cytotoxic t-cell response in women identified by microarray analysis. *Genes & Immunity* **10**, 509–516 (2009). URL <https://doi.org/10.1038/gene.2009.12>.
- [63] Klein, S. L. & Flanagan, K. L. Sex differences in immune responses. *Nature Reviews Immunology* **16**, 626–638 (2016). URL <https://doi.org/10.1038/nri.2016.90>.
- [64] Takahashi, T. *et al.* Sex differences in immune responses that underlie COVID-19 disease outcomes. *Nature* (2020). URL <https://doi.org/10.1038/s41586-020-2700-3>.
- [65] Linton, P. J. & Dorshkind, K. Age-related changes in lymphocyte development and function. *Nature Immunology* **5**, 133–139 (2004). URL <https://doi.org/10.1038/ni1033>.
- [66] Elyahu, Y. *et al.* Aging promotes reorganization of the cd4 t cell landscape toward extreme regulatory and effector phenotypes. *Science Advances* **5** (2019). URL <https://advances.sciencemag.org/content/5/8/eaaw8330>. <https://advances.sciencemag.org/content/5/8/eaaw8330.full.pdf>.
- [67] McPadden, J. *et al.* Clinical characteristics and outcomes for 7,995 patients with sars-cov-2 infection. *medRxiv* (2020). URL <https://www.medrxiv.org/content/early/2020/07/21/2020.07.19.20157305>. <https://www.medrxiv.org/content/early/2020/07/21/2020.07.19.20157305.full.pdf>.
- [68] Williamson, E. J. *et al.* Factors associated with COVID-19-related death using OpenSAFELY. *Nature* **584**, 430–436 (2020). URL <https://doi.org/10.1038/s41586-020-2521-4>.
- [69] Chen, Z. & John Wherry, E. T cell responses in patients with COVID-19. *Nat Rev Immunol* **20**, 529–536 (2020).

- [70] Sekine, T. *et al.* Robust T Cell Immunity in Convalescent Individuals with Asymptomatic or Mild COVID-19. *Cell* **183**, 158–168 (2020).
- [71] Zhang, F. *et al.* Adaptive immune responses to SARS-CoV-2 infection in severe versus mild individuals. *Signal Transduct Target Ther* **5**, 156 (2020).
- [72] Sattler, A. *et al.* SARS-CoV-2 specific T-cell responses and correlations with COVID-19 patient predisposition. *J Clin Invest* (2020).
- [73] Odak, I. *et al.* Reappearance of effector T cells is associated with recovery from COVID-19. *EBioMedicine* **57**, 102885 (2020).
- [74] Coifman, R. R. & Lafon, S. Diffusion maps. *Applied and Computational Harmonic Analysis* **21**, 5–30 (2006).
- [75] Bermanis, A., Wolf, G. & Averbuch, A. Diffusion-based kernel methods on euclidean metric measure spaces. *Applied and Computational Harmonic Analysis* – (2015). URL <http://www.sciencedirect.com/science/article/pii/S1063520315001013>.
- [76] Moon, K. R. *et al.* Manifold learning-based methods for analyzing single-cell RNA-sequencing data. *Current Opinion in Systems Biology* **7**, 36–46 (2018).
- [77] Bermanis, A., Wolf, G. & Averbuch, A. Cover-based bounds on the numerical rank of gaussian kernels. *Applied and Computational Harmonic Analysis* **36**, 302 – 315 (2014).
- [78] Belkin, M. & Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation* **15**, 1373–1396 (2003).
- [79] Gigante, S., Charles, A. S., Krishnaswamy, S. & Mishne, G. Visualizing the phate of neural networks (2019). [1908.02831](https://arxiv.org/abs/1908.02831).
- [80] David, G. & Averbuch, A. Hierarchical data organization, clustering and denoising via localized diffusion folders. *Applied and Computational Harmonic Analysis* **33**, 1–23 (2012).
- [81] Wolf, G., Rotbart, A., David, G. & Averbuch, A. Coarse-grained localized diffusion. *Applied and Computational Harmonic Analysis* **33**, 388–400 (2012).
- [82] Vietoris, L. Über den höheren Zusammenhang kompakter Räume und eine Klasse von zusammenhangstreuen Abbildungen. *Mathematische Annalen* **97**, 454–472 (1927).
- [83] Marshall, N. F. & Hirn, M. J. Time coupled diffusion maps. *Applied and Computational Harmonic Analysis* **45**, 709–728 (2018).
- [84] Gigante, S. *et al.* Compressed diffusion. In *The 13th International Conference on Sampling Theory and Applications (SampTA 2019)* (Bordeaux, France, 2019).
- [85] Savaresi, S. M. & Boley, D. L. On the performance of bisecting k-means and pddp. In *Proceedings of the 2001 SIAM International Conference on Data Mining*, 1–14 (SIAM, 2001).
- [86] Grygorash, O., Zhou, Y. & Jorgensen, Z. Minimum spanning tree based clustering algorithms. In *2006 18th IEEE International Conference on Tools with Artificial Intelligence (ICTAI'06)*, 73–81 (IEEE, 2006).

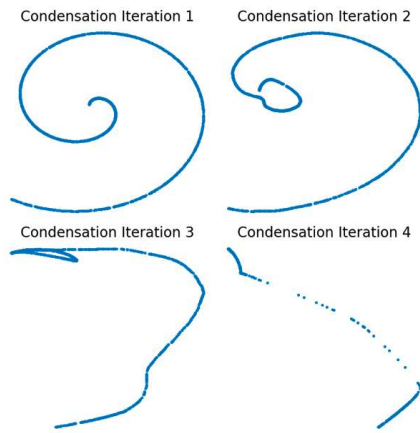
- [87] Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring network structure, dynamics, and function using networkx. In Varoquaux, G., Vaught, T. & Millman, J. (eds.) *Proceedings of the 7th Python in Science Conference*, 11 – 15 (Pasadena, CA USA, 2008).
- [88] Pedregosa, F. *et al.* Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* **12**, 2825–2830 (2011).
- [89] Chen, W. S. *et al.* Uncovering axes of variation among single-cell cancer specimens. *Nat Methods* **17**, 302–310 (2020).
- [90] Leeb, W. & Coifman, R. Hölder–lipschitz norms and their duals on spaces with semigroups, with applications to earth mover’s distance. *Journal of Fourier Analysis and Applications* **22**, 910–953 (2015). URL <https://doi.org/10.1007/s00041-015-9439-5>.
- [91] Cusanovich, D. A. *et al.* A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309–1324.e18 (2018). URL <https://doi.org/10.1016/j.cell.2018.06.052>.
- [92] Hartmann, F. J. *et al.* Comprehensive immune monitoring of clinical trials to advance human immunotherapy. *Cell Reports* **28**, 819–831.e4 (2019). URL <https://doi.org/10.1016/j.celrep.2019.06.049>.
- [93] Nolan, S. *et al.* A large-scale database of t-cell receptor beta (TCR $\beta$ ) sequences and binding associations from natural and synthetic exposure to SARS-CoV-2. (2020). URL <https://doi.org/10.21203/rs.3.rs-51964/v1>.
- [94] Corrie, B. D. *et al.* iReceptor: A platform for querying and analyzing antibody/b-cell and t-cell receptor repertoire data across federated repositories. *Immunological Reviews* **284**, 24–41 (2018). URL <https://doi.org/10.1111/imr.12666>.



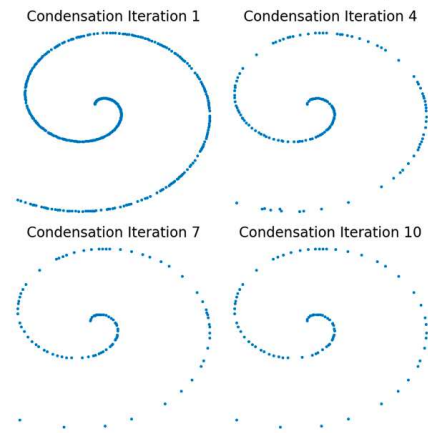
## 9 Extended Data Figures

**a**

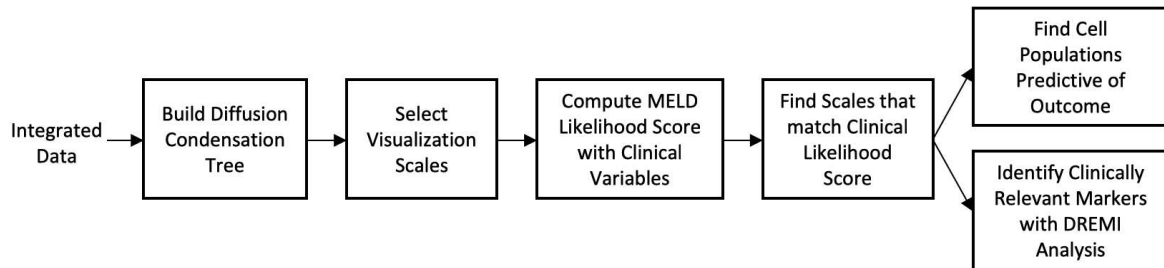
**Multiscale PHATE in Euclidean Space**



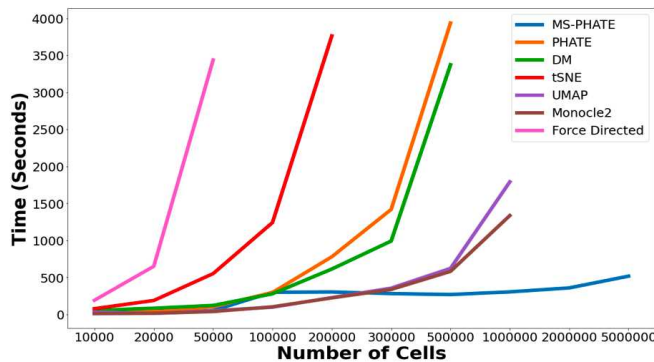
**Multiscale PHATE in Diffusion Space**



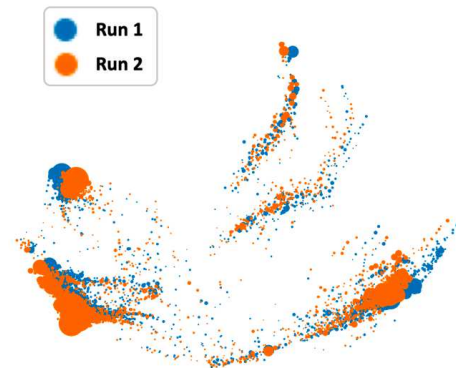
**b**



**c**



**d**



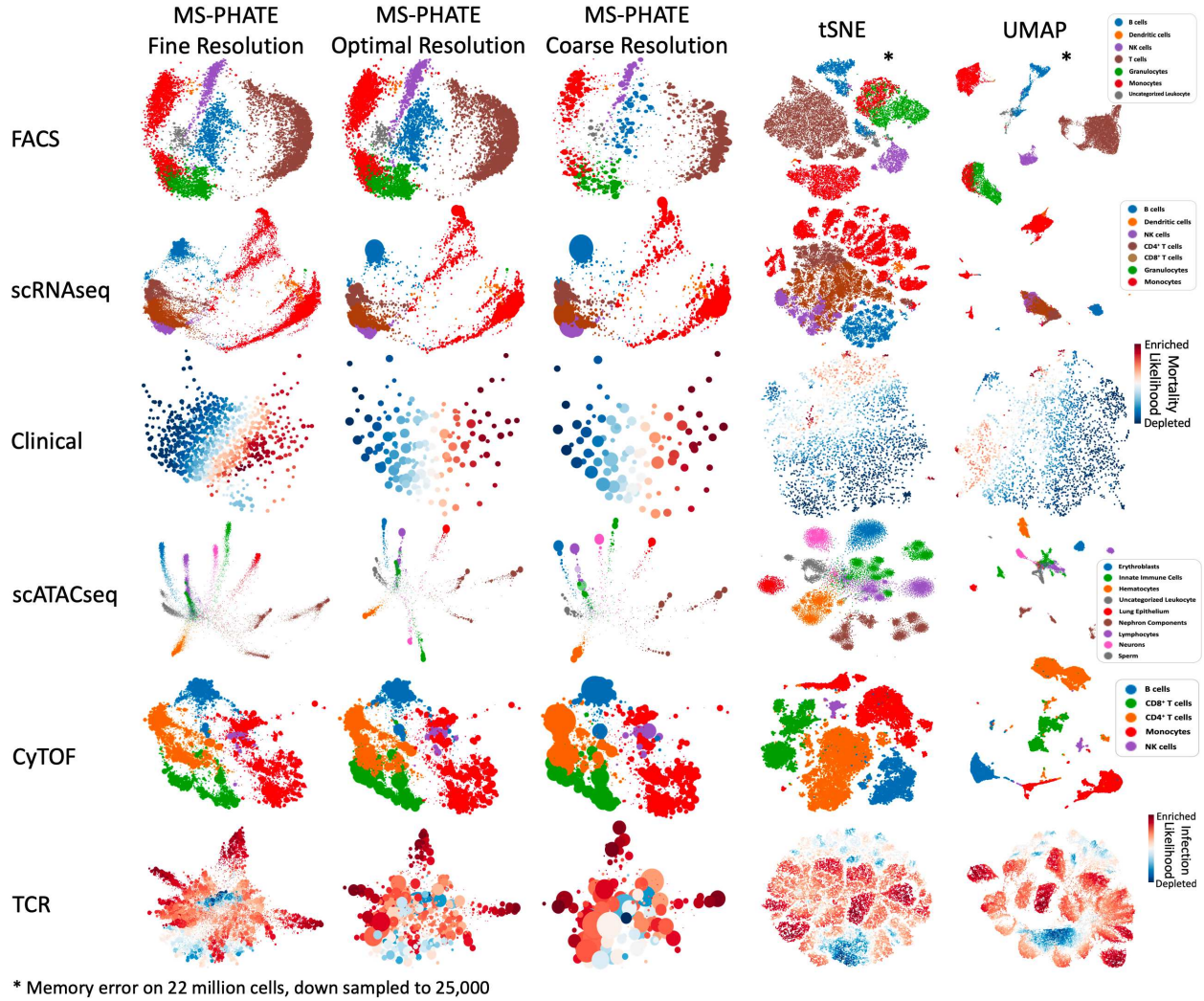
**Extended Data Figure 1: Condensing on Manifold and updating visualization.**

*a. Visualization of toy swiss roll dataset after several iterations of fast diffusion condensation, running in both feature space and in manifold space as computed by Diffusion Potential.*

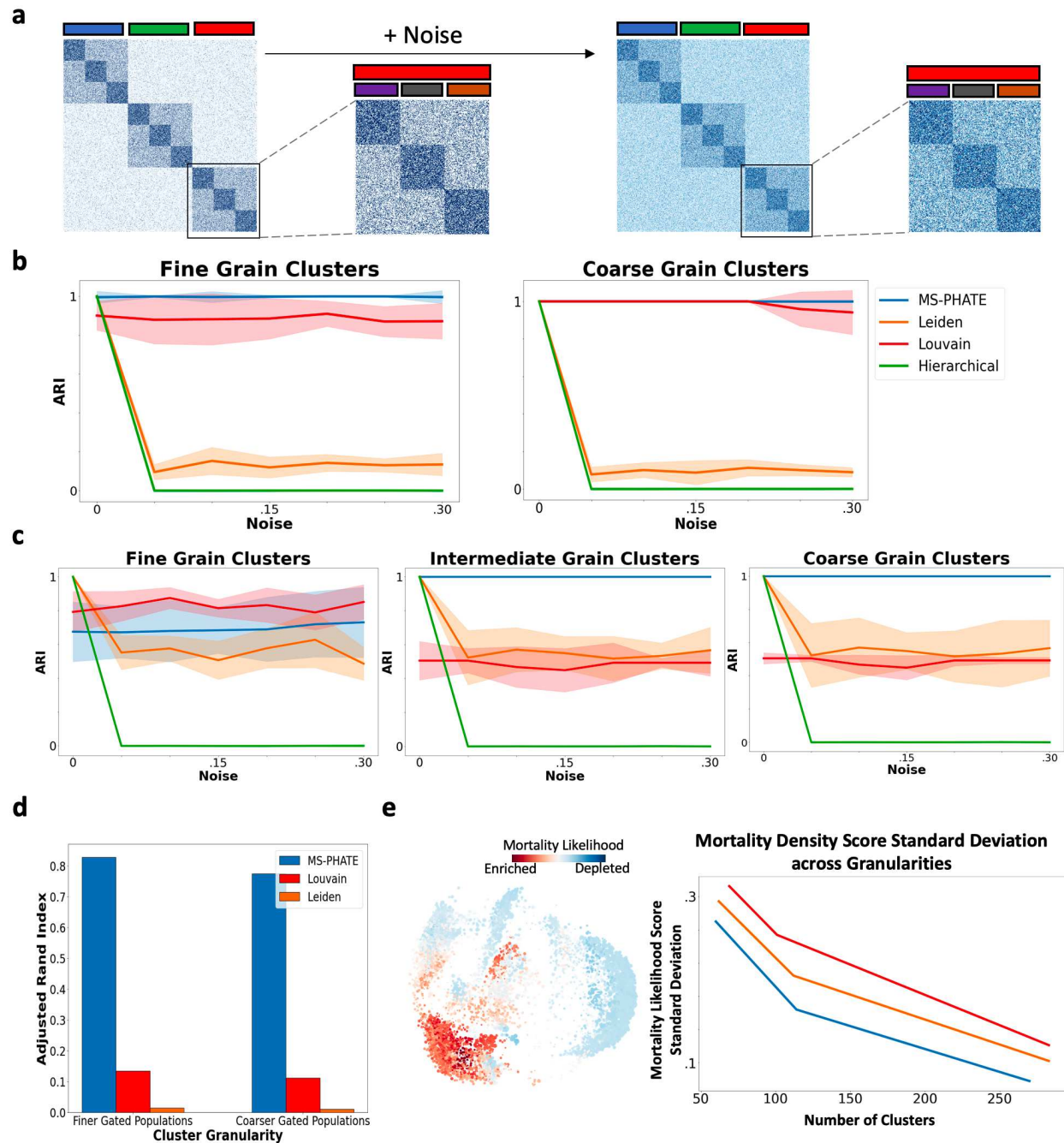
*b. Pipeline for identifying cellular populations enriched based on clinical variables with Multiscale PHATE and MELD.*

*c. Comparing run time across visualization techniques on increasingly high dimensional flow cytometry data.*

*d. Visualization of reproducibility of Multiscale PHATE across two different runs of PBMCs measured by scRNAseq. Each run was initialized with a different random seed.*

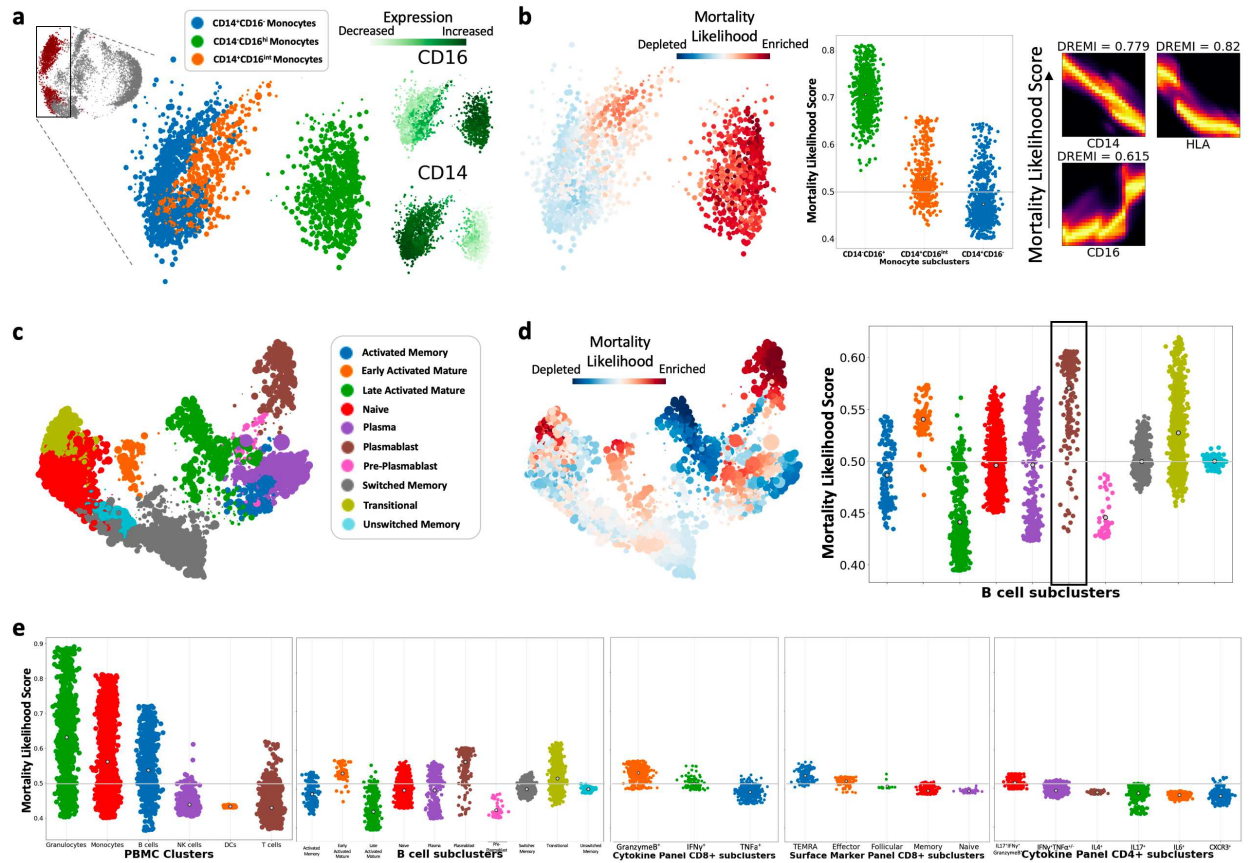


**Extended Data Figure 2: Visualization of differing high dimensional biological data types.** Visualization comparison across a range of data types: 22 million PBMCs measured by flow cytometry [14], 49,942 PBMCs by scRNAseq [10], 2,135 patients admitted to YNHH by demographic and lab clinical variables, 25,528 cells from a diverse set of mouse tissues measured by scATACseq [91], 1,010,964 PBMCs measured by CyTOF [92] and 50,000 TCRs from COVID-19 infected patients and healthy controls [93, 94].



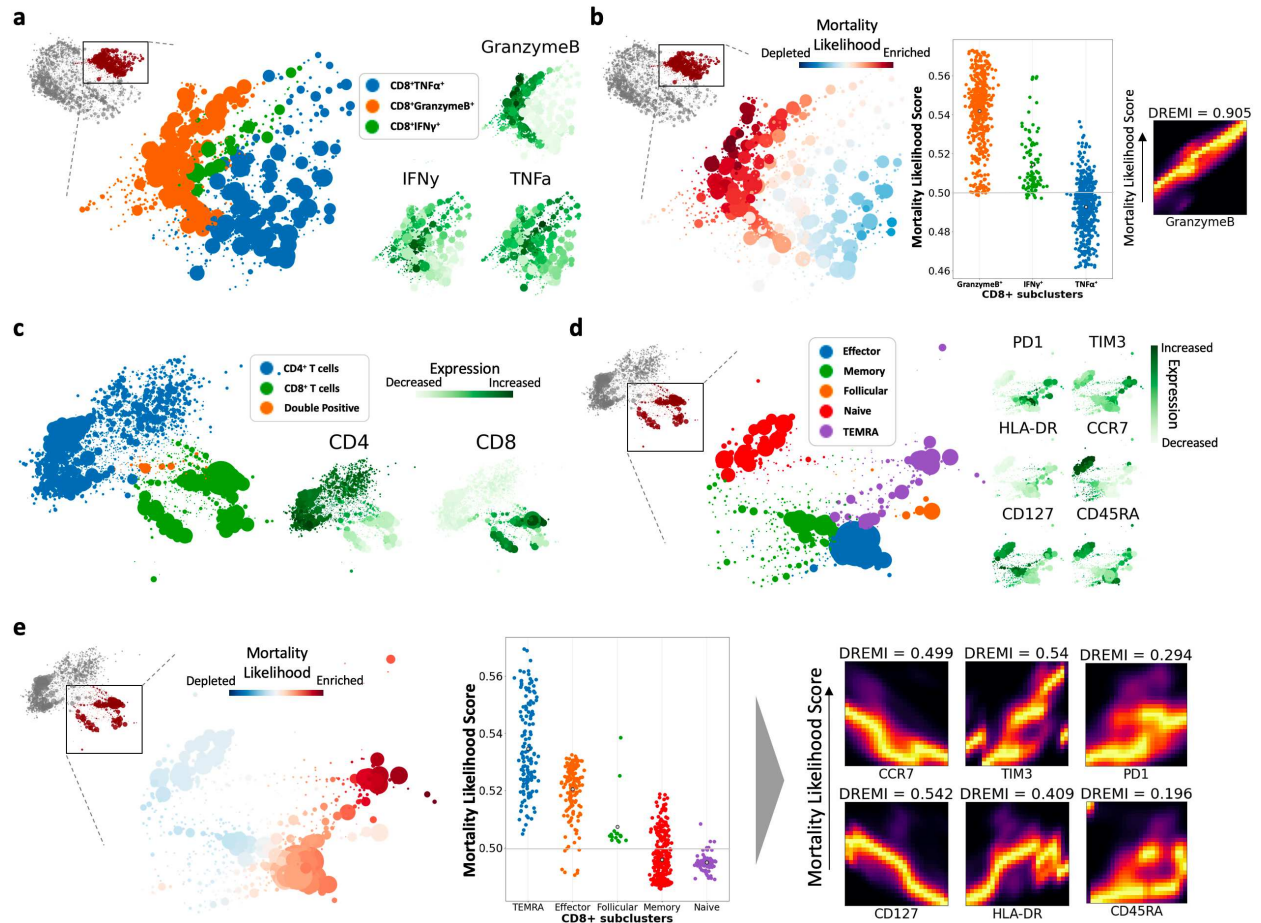
**Extended Data Figure 3: Comparison of Multiscale PHATE with other Clustering techniques.**  
*a.* Schematic of the hierarchical stochastic block model we generated for multigranular cluster comparisons. For each method, increasing amounts of random Gaussian values were added to the adjacency matrix of stochastic block model to simulate increasing amounts of noise.  
*b.* Computed Adjusted Rand Index (ARI) between each algorithm's predicted clusters and the known clusters across coarse and fine granularities of 2 layer stochastic block model.  
*c.* Computed Adjusted Rand Index (ARI) between each algorithm's predicted clusters and the known clusters across coarse and fine granularities of 3 layer stochastic block model.  
*d.* Comparison of multiple clustering approaches on flow cytometry data where cell types and subtypes have been identified through gating analysis. Clusters identified by different approaches were compared to gated populations using ARI.  
*e.* Comparison of multiple clustering techniques at identifying regions with uniform MELD likelihood scores across a range of comparable granularities.





**Extended Data Figure 4: Multiscale PHATE identifies subsets of monocytes and B cells enriched in patients who died from COVID-19.**

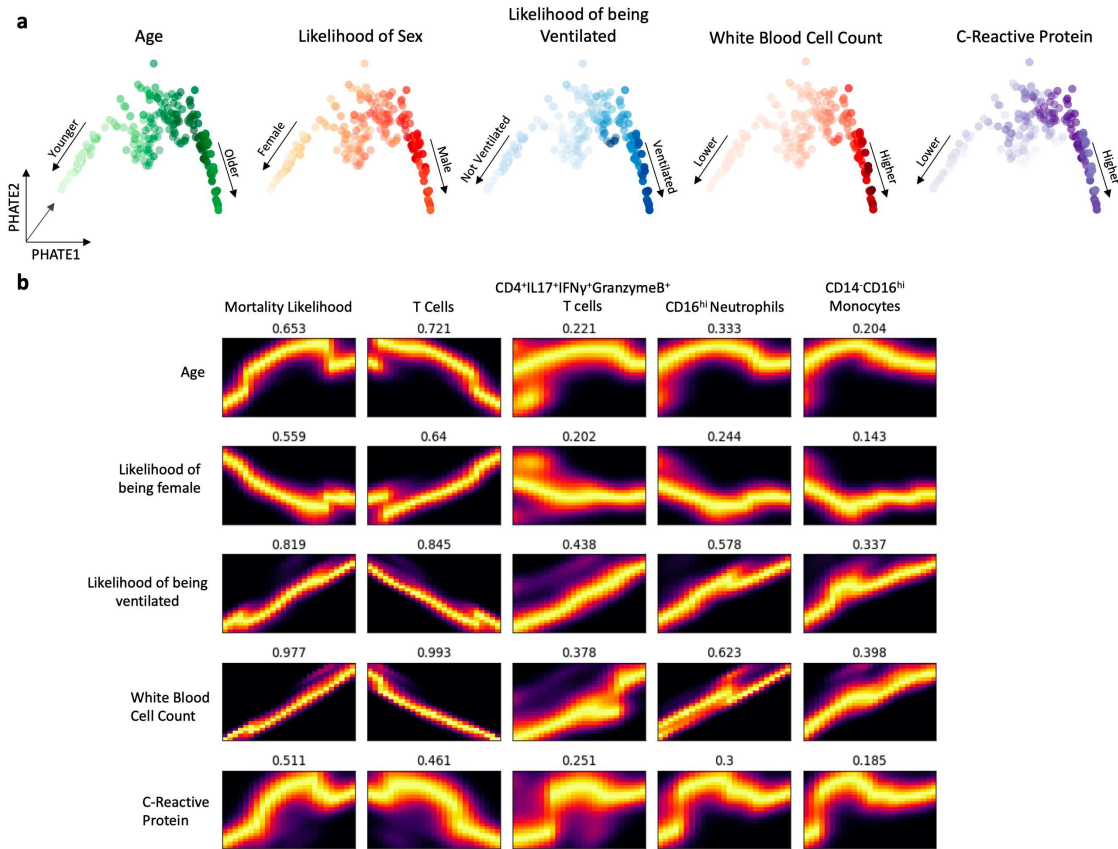
- Zoom in of monocyte population identifies subsets based on expression of markers.
- Visualization of mortality likelihood score in monocytes identifies subsets enriched in patients who die from COVID-19. Key associations between markers and mortality likelihood score computed by DREMI and visualized with DREVI.
- Visualization of B cells panel identifies a range of subsets based on expression of known markers.
- Visualization of mortality likelihood score identifies B cell subsets enriched in patients who die from COVID-19.
- Comparison of mortality likelihood score across panels reveals that granulocytes and monocytes are broadly the most enriched cell types in patients who die from COVID-19.



**Extended Data Figure 5: Multiscale PHATE analysis identifies subsets of CD8<sup>+</sup> T cells enriched in patients with poor COVID-19 outcomes.**

- Zoom in of CD8<sup>+</sup> T cells identifies subsets based on expression of markers.
- Visualization of mortality likelihood score in CD8<sup>+</sup> T cells identifies subsets enriched in patients who die from COVID-19. Key associations between GranzymeB and mortality likelihood computed by DREMI and visualized with DREVI.
- Multiscale PHATE visualization of T cell focused surface marker panel with broad T cell subtypes identified.
- Zoom in of CD8<sup>+</sup> T cells identifies subsets based on expression of known markers.
- Visualization of mortality likelihood score in CD8<sup>+</sup> T cells identifies subsets enriched in patients who die from COVID-19. Key associations between markers and mortality likelihood computed by DREMI and visualized with DREVI.

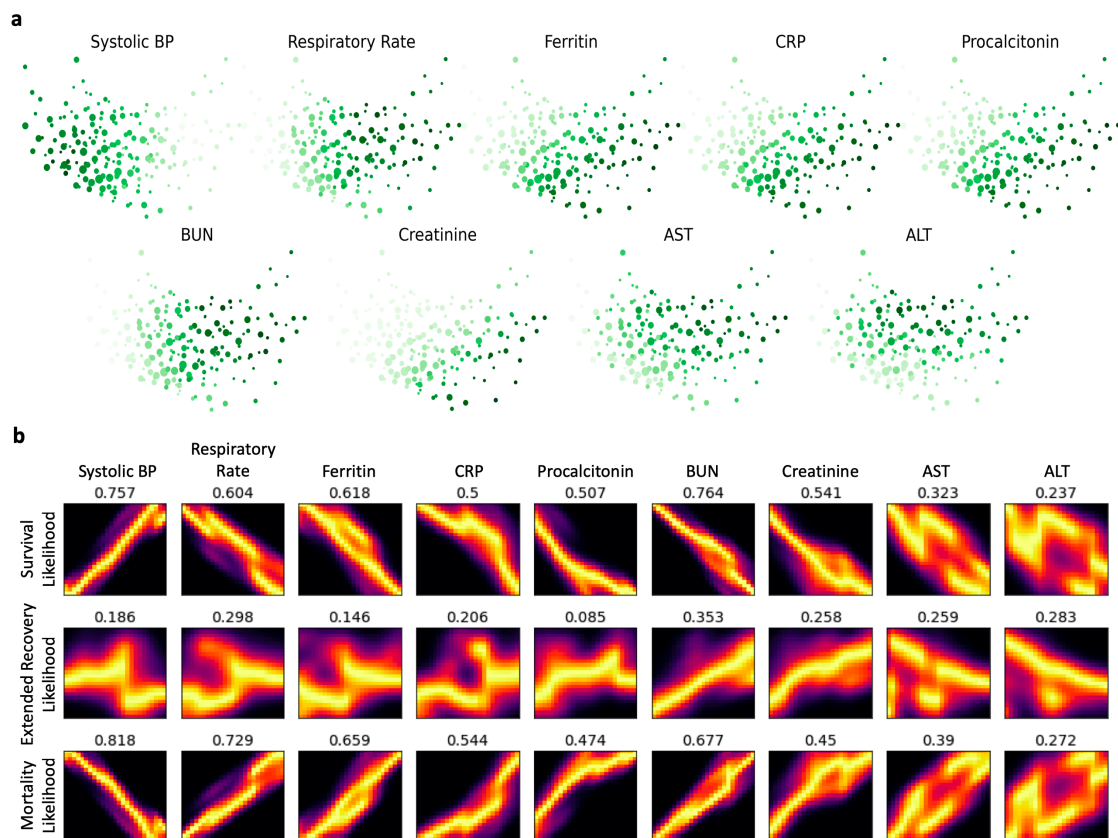




**Extended Data Figure 6: Visualization of patient manifold and correlation with clinical features**

*a. Visualizing clinical trends on patient manifold.*

*b. DREMI and DREVI association analysis between clinical features and mortality as well as cellular populations.*

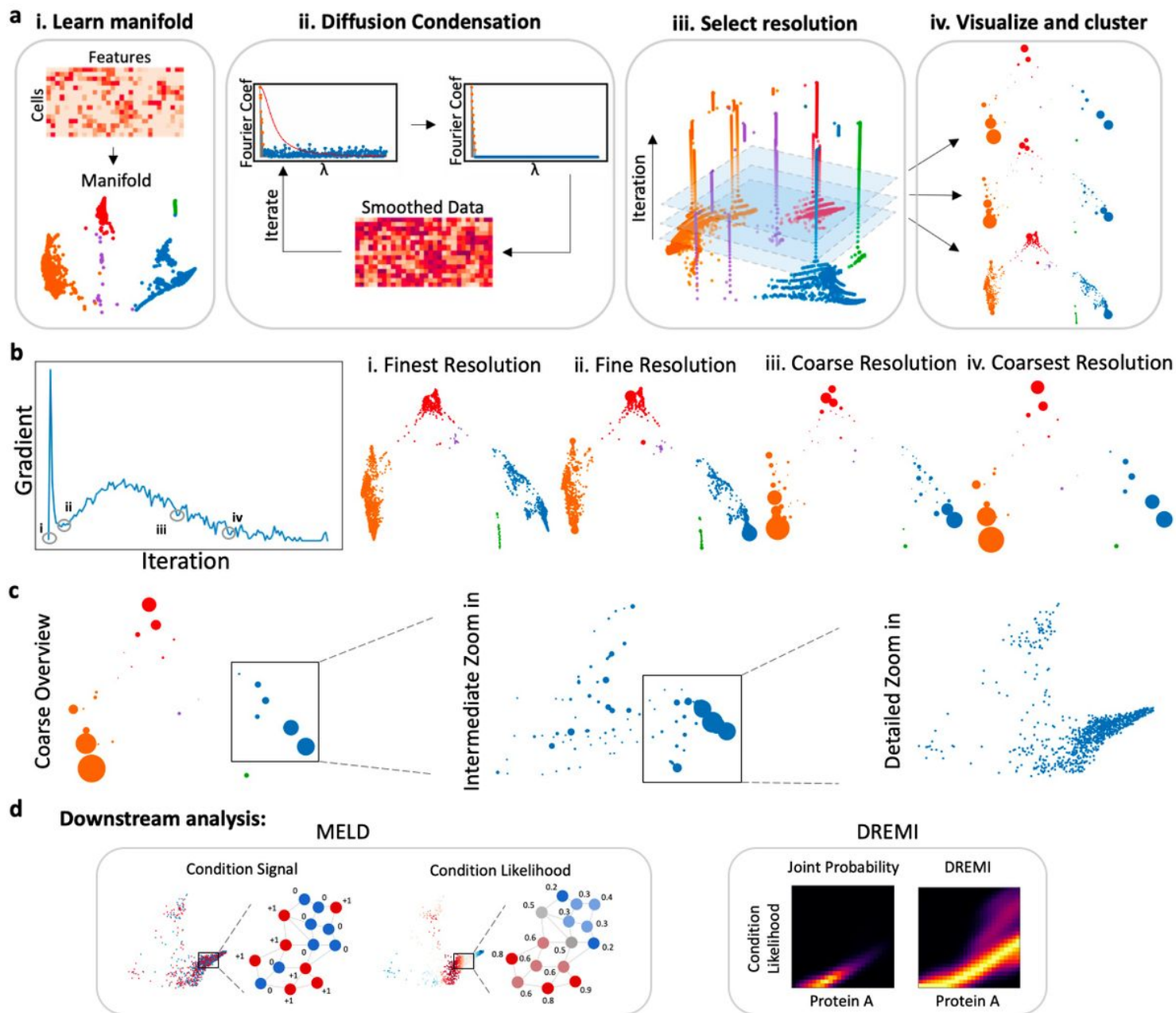


**Extended Data Figure 7: Visualization of multiscale clinical manifold and correlation with patient clinical features.**

*a. Visualizing clinical trends on clinical manifold. Darker color indicates higher normalized numerical values.*

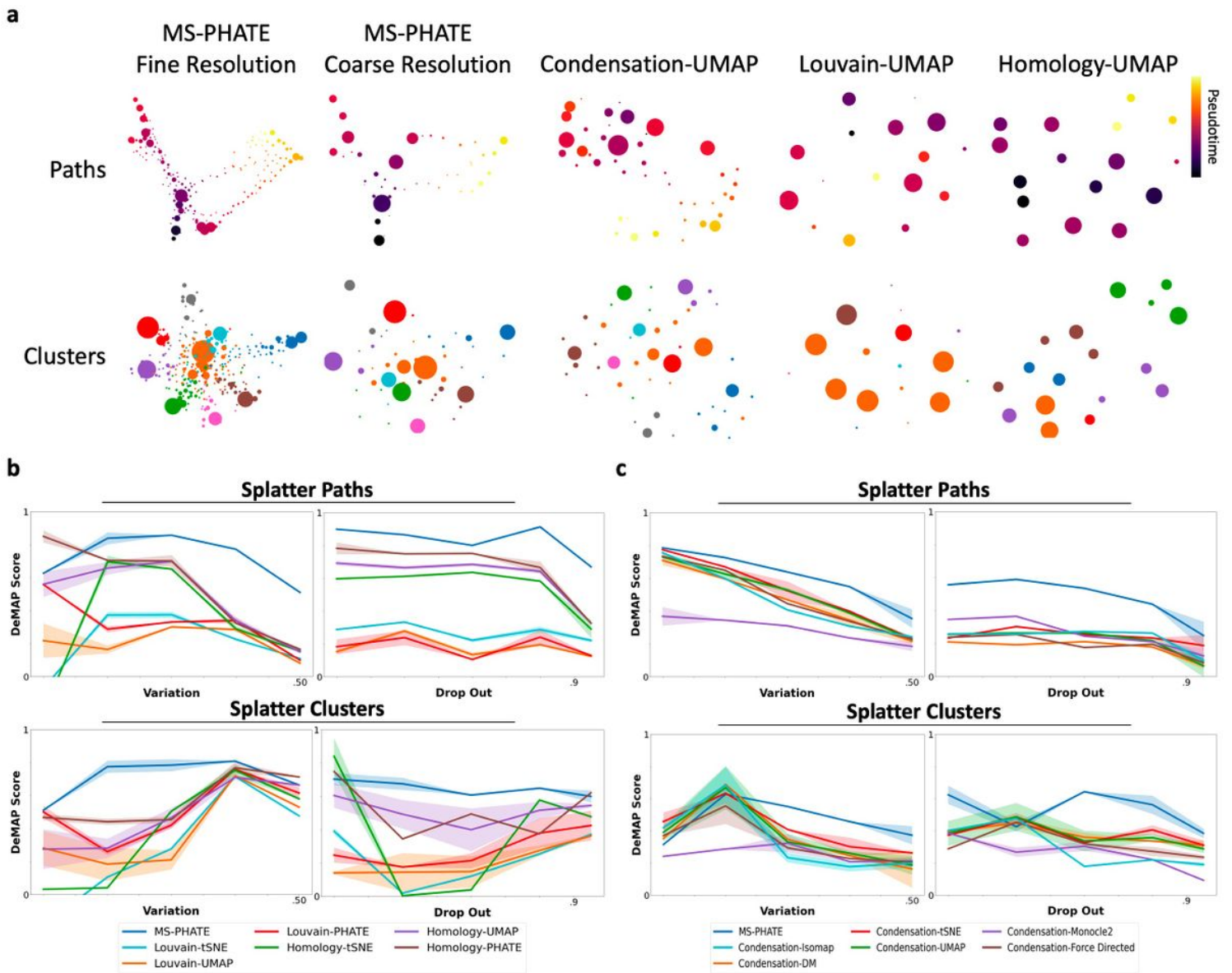
*b. DREMI and DREVI association analysis between clinical features and patient hospitalization outcome.*

# Figures



**Figure 1**

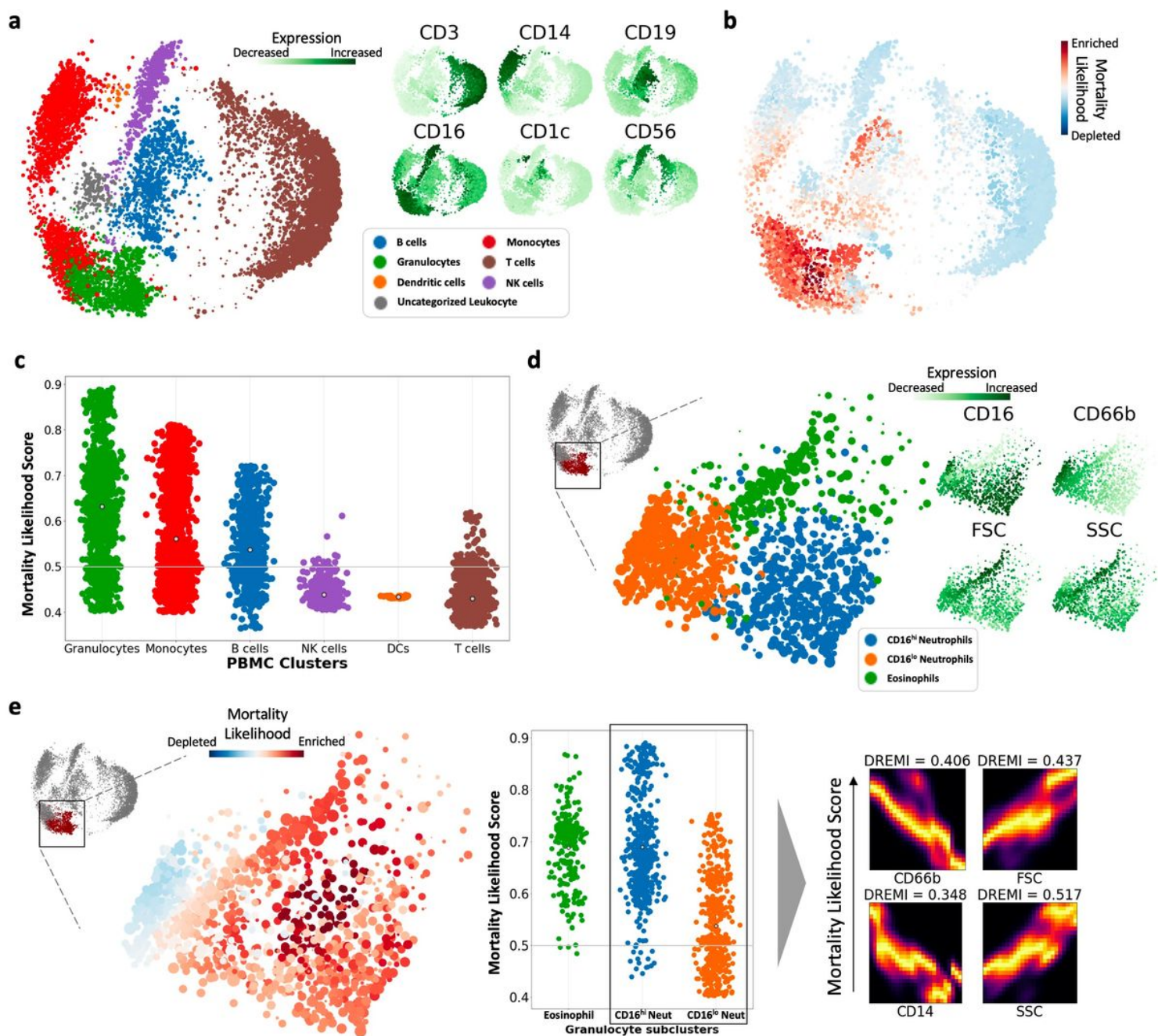
Overview of Multiscale PHATE algorithm a. Multiscale PHATE process involves four successive steps. The first step (i) learns the manifold geometry via diffusion potential calculation. The second step (ii) iteratively coarse grains the manifold construction through a fast diffusion condensation process to learn data topology. The third step (iii) involves the selection of salient granularities via gradient analysis before finally visualizing and clustering the manifold in the fourth step (iv). b. Gradient analysis identifies a range of scales for visualization. c. Multiscale PHATE allows for high level summarizations of data as well as finer grain zoom ins of data subsets for additional detail. d. Multiscale PHATE abstractions of data are amenable to downstream analyses with algorithms like MELD [15] and DREMI [16].



**Figure 2**

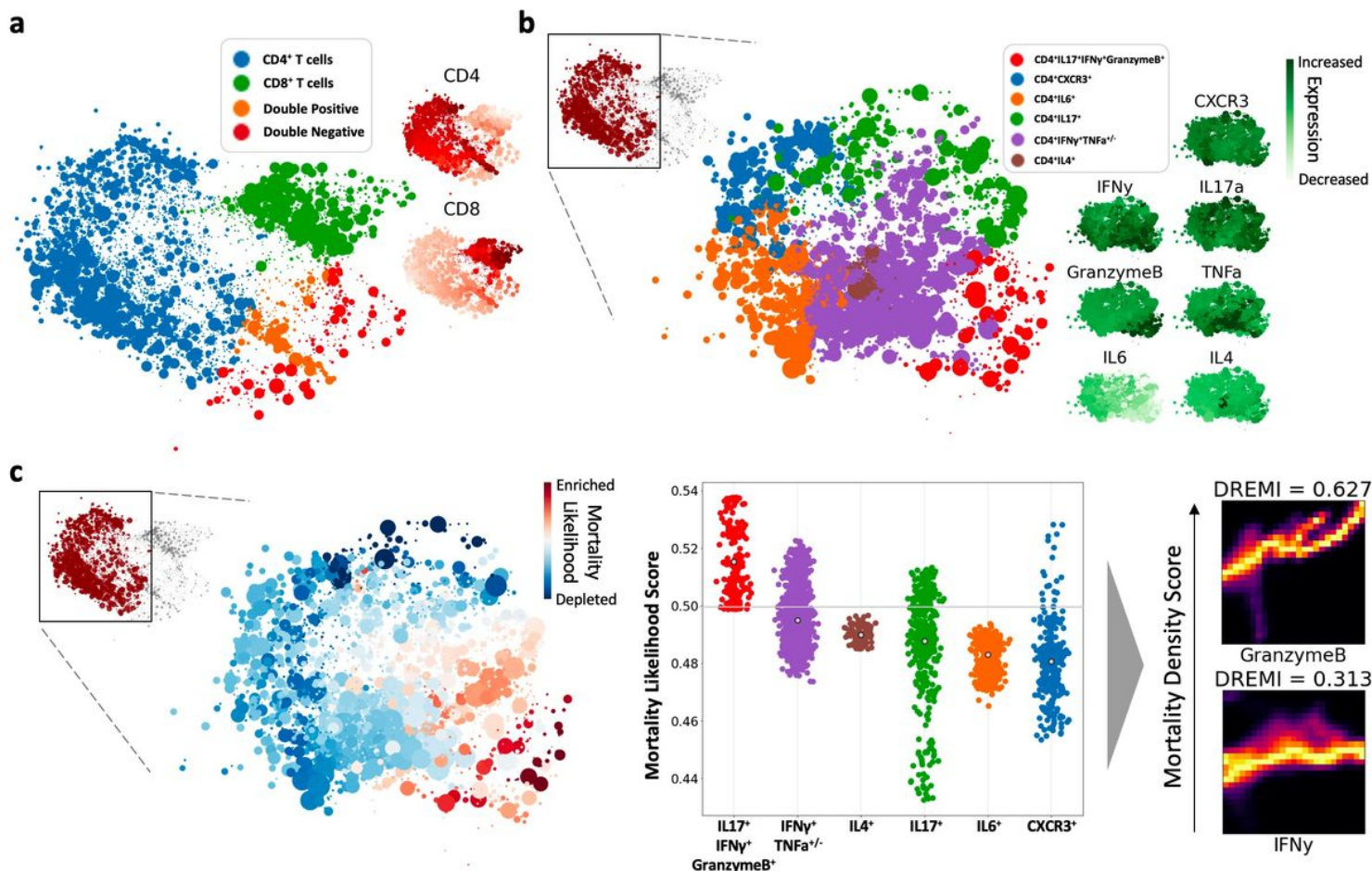
Comparison of Multiscale PHATE with other dimensionality reduction tools a. Visual comparison of Multiscale PHATE with other multiscale dimensionality reduction tools on synthetic single cell data with either path or cluster structure. b. Quantitative study comparing embeddings produced by Multiscale PHATE and visualization strategies which either employ community based or topologically based abstractions of data. Comparisons were evaluated using DeMAP with increasing levels of 2 different types of biological noise, drop out and variation, as well as on data with different structures, clusters and paths. Shading represents standard deviation around mean DeMAP score for each comparison. c. Quantitative study comparing embeddings produced by Multiscale PHATE and visualization strategies which visualize condensation based abstractions of data. Comparisons were evaluated using DeMAP with increasing levels of 2 different types of biological noise, drop out and variation, as well as on data with different structures, clusters and paths. Shading represents standard deviation around mean DeMAP score for each comparison.





**Figure 3**

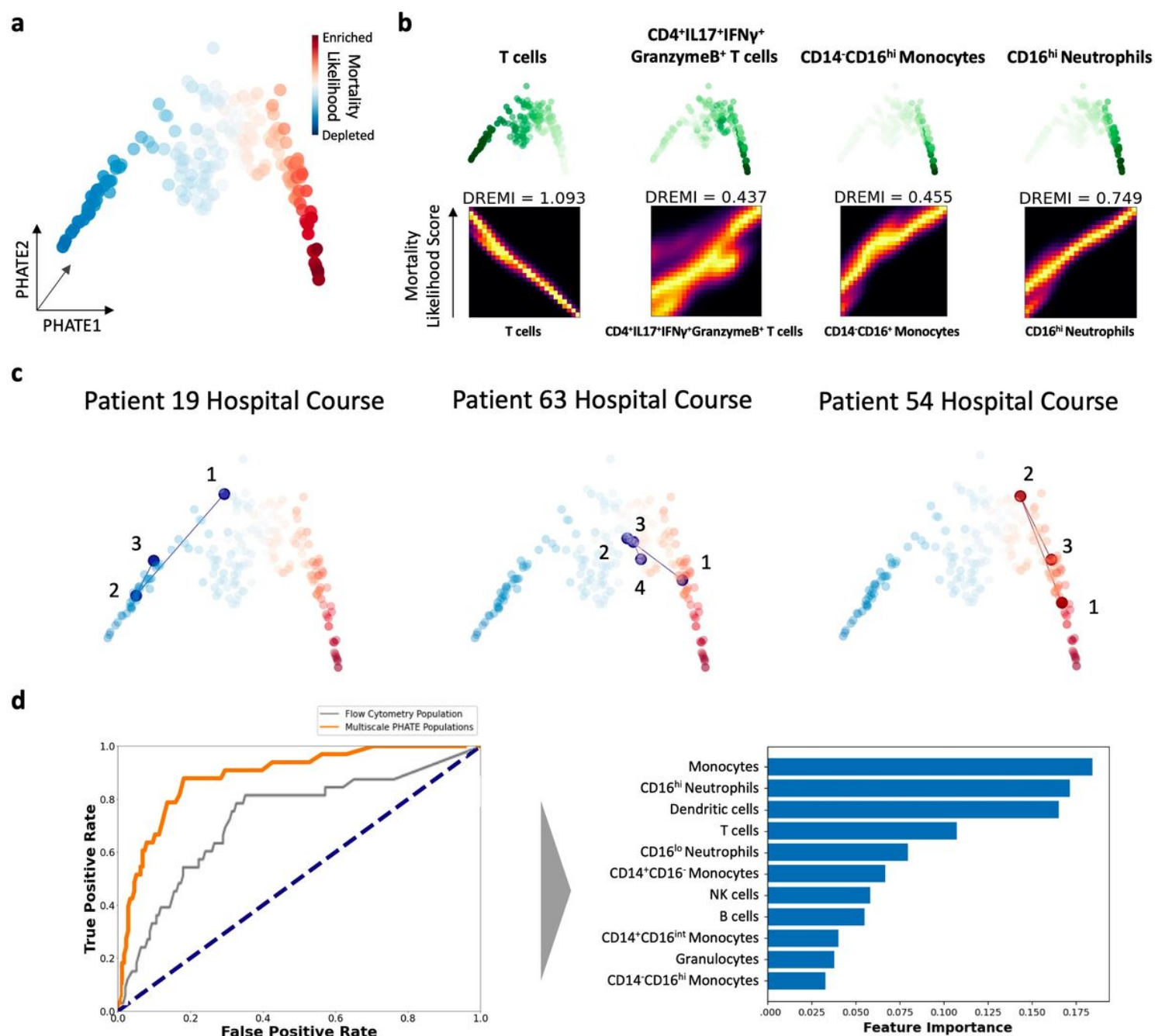
CD16<sup>hi</sup>CD66b<sup>lo</sup> Neutrophil subset enriched in patients who die from COVID-19. a. Multiscale PHATE visualization of PBMCs identifies all major cell types based on cell type specific markers. b. Visualization of mortality likelihood score computed by MELD. c. Visualization of mortality likelihood score organized by cell type reveals enrichment of granulocytes, monocytes and B cells in patients who die from COVID-19. d. Zoom in of granulocyte population identifies subsets of neutrophils and eosinophils based on expression of known markers. e. Visualization of mortality likelihood score in granulocyte population identifies CD16<sup>hi</sup> neutrophils enriched in patients with worse outcomes. Key associations between markers and mortality likelihood scores in neutrophils computed by DREMI and visualized with DREVI.



**Figure 4**

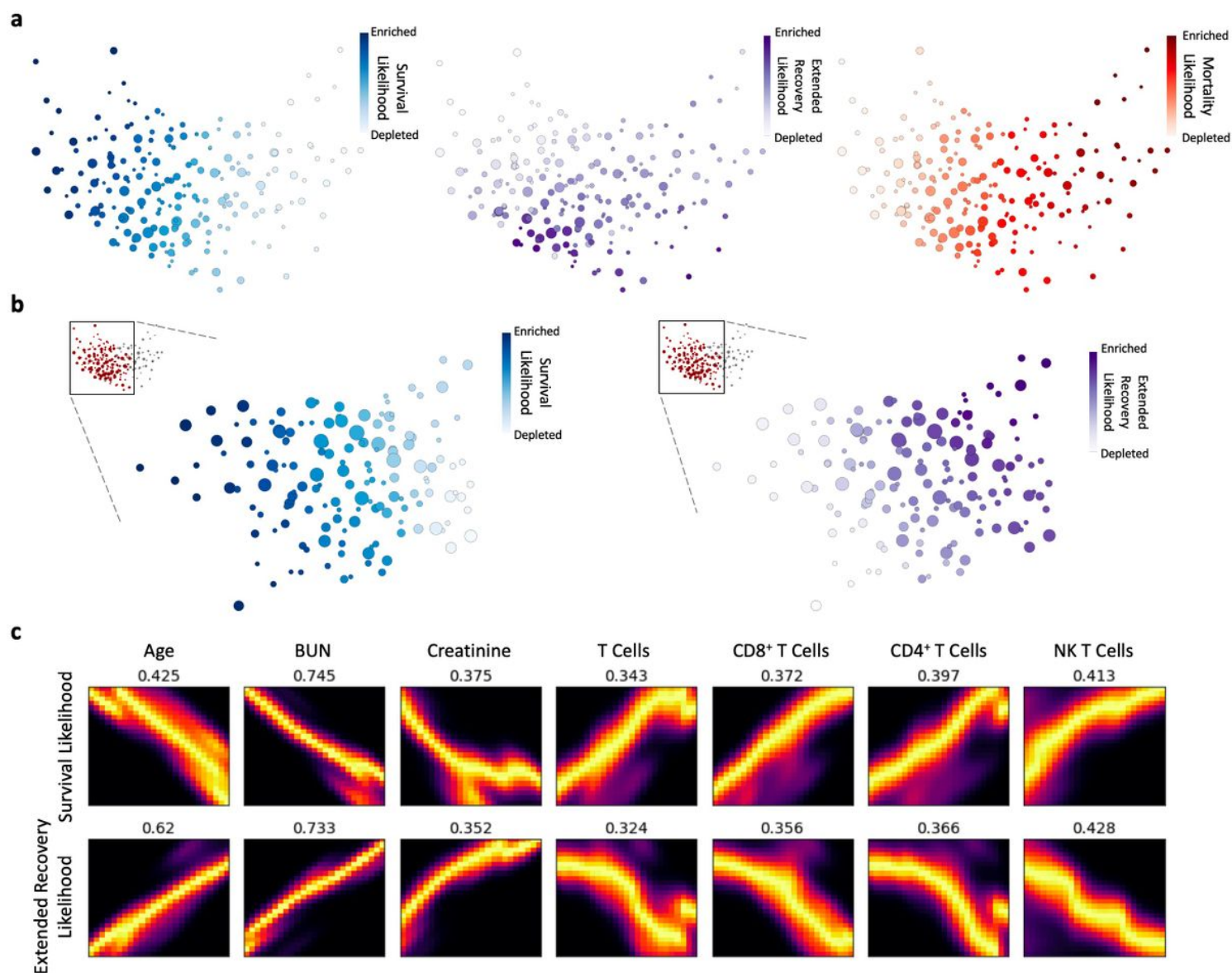
Multiscale PHATE identifies Th17 subset enriched in patients who die from COVID-19 a. Multiscale PHATE visualization of T cell focused cytokine panel identifies broad T cell subtypes. b. Zoom in of CD4<sup>+</sup> T helper cells identifies subsets based on expression of functional markers. c. Visualization of mortality likelihood score identifies IFN $\gamma$ +GranzymeB<sup>+</sup> Th17 cell enrichment in patients with poor outcomes. Key associations between markers and mortality likelihood scores are computed by DREMI and visualized with DREVI.





**Figure 5**

Patient manifold corroborates cellular states associated with disease pathogenesis. a. Visualization of patient manifold via PHATE and mortality likelihood score based on patient outcomes computed via MELD. b. Visualization of key cell population enrichment trends over the manifold with associations computed by DREMI and visualized with DREVI. c. Tracing three patients' hospital courses over patient manifold. Patients 19 and 63 were discharged while patient 54 died. d. Comparing predictability of patient mortality using random forest classifier on Multiscale PHATE identified populations and flow cytometry identified populations. Most predictive Multiscale PHATE clusters are ranked through feature importance analysis.



**Figure 6**

Multiscale manifold of patient clinical features identifies cell types associated with extended COVID-19 recovery phase a. Visualization of Multiscale PHATE clinical manifold constructed on patient clinical features. Embedding is colored by likelihood scores based on patient outcomes computed via MELD. b. Zoom in on transition point between high extended recovery likelihood score and high survival likelihood score. c. Patient clinical features and flow cytometry identified cell populations associated with patient outcomes using DREMI and visualized with DREVI.