nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-021-01186-x

Check for updates

Multiscale PHATE identifies multimodal signatures of COVID-19

Manik Kuchroo[1,30], Jessie Huang[2,30], Patrick Wong[3,30], Jean-Christophe Grenier[4], Dennis Shung[5], Alexander Tong[2], Carolina Lucas[3], Jon Klein[3], Daniel B. Burkhardt[6], Scott Gigante[7], Abhinav Godavarthi[8], Bastian Rieck[9], Benjamin Israelow[3,10], Michael Simonov[5], Tianyang Mao[3], Ji Eun Oh[3], Julio Silva[3], Takehiro Takahashi[3], Camila D. Odio[5], Arnau Casanovas-Massana[11], John Fournier[10], Yale IMPACT Team*, Shelli Farhadian[10], Charles S. Dela Cruz[12,13], Albert I. Ko[10,11], Matthew J. Hirn[14,15], F. Perry Wilson[16], Julie G. Hussin[4,17,31], Guy Wolf[18,19,31], Akiko Iwasaki[3,20,31] and Smita Krishnaswamy[2,6,31] ✉

As the biomedical community produces datasets that are increasingly complex and high dimensional, there is a need for more sophisticated computational tools to extract biological insights. We present Multiscale PHATE, a method that sweeps through all levels of data granularity to learn abstracted biological features directly predictive of disease outcome. Built on a coarse-graining process called diffusion condensation, Multiscale PHATE learns a data topology that can be analyzed at coarse resolutions for high-level summarizations of data and at fine resolutions for detailed representations of subsets. We apply Multiscale PHATE to a coronavirus disease 2019 (COVID-19) dataset with 54 million cells from 168 hospitalized patients and find that patients who die show CD16^{hi}CD66b^{lo} neutrophil and IFN-γ^{+} granzyme B^{+} Th17 cell responses. We also show that population groupings from Multiscale PHATE directly fed into a classifier predict disease outcome more accurately than naive featurizations of the data. Multiscale PHATE is broadly generalizable to different data types, including flow cytometry, single-cell RNA sequencing (scRNA-seq), single-cell sequencing assay for transposase-accessible chromatin (scATAC-seq), and clinical variables.

High-throughput biomedical data are generated by a range of technologies[1–3] that measure dozens to tens of thousands of features in millions of cells derived from large patient cohorts. We posit here that the key to understanding such 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 (*t*-SNE)[4], uniform manifold approximation and projection (UMAP)[5] and principal-component analysis (PCA)[6], only show a single level of granularity of the data. Recent papers on severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) (refs. [7,8]) have used one of these approaches to understand patient cellular responses at a single resolution. Differences between an effective immunological response and an ineffective one, however, may not be found at the granularity of immune compartment abundance alone.

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 topological process called diffusion condensation[9], which slowly condenses data points toward local centers of gravity to form natural, data-driven groupings across granularities. This coarse-graining process continuously learns the topology of the underlying dataset by allowing cells to naturally come together over the course of successive condensation steps, allowing for the exploration of a more continuous range of granularities not revealed through other methods. Visualizing a series of

[1]Department of Neuroscience, Yale University, New Haven, CT, USA. [2]Department of Computer Science, Yale University, New Haven, CT, USA. [3]Department of Immunobiology, Yale University, New Haven, CT, USA. [4]Montreal Heart Institute, Montreal, Quebec, Canada. [5]Department of Medicine, Yale University, New Haven, CT, USA. [6]Department of Genetics, Yale University, New Haven, CT, USA. [7]Computational Biology, Bioinformatics Program, Yale University, New Haven, CT, USA. [8]Department of Applied Mathematics, Yale University, New Haven, CT, USA. [9]Department of Biosystems Science and Engineering, ETH Zurich, Zurich, Switzerland. [10]Section of Infectious Diseases, Department of Medicine, Yale University School of Medicine, New Haven, CT, USA. [11]Department of Epidemiology of Microbial Diseases, Yale School of Public Health, New Haven, CT, USA. [12]Section of Pulmonary and Critical Care Medicine, Department of Medicine, Yale University School of Medicine, New Haven, CT, USA. [13]Department of Medicine, West Haven Connecticut Veterans Affairs Medical Center, West Haven, CT, USA. [14]Department of Mathematics, Michigan State University, East Lansing, MI, USA. [15]Department of Computational Mathematics, Science and Engineering, Michigan State University, East Lansing, MI, USA. [16]Clinical and Translational Research Accelerator, Department of Medicine, Yale University, New Haven, CT, USA. [17]Faculty of Medicine, Université de Montréal, Montreal, Quebec, Canada. [18]Mila – Quebec AI institute, Montreal, Quebec, Canada. [19]Department of Mathematics and Statistics, Université de Montréal, Montreal, Quebec, Canada. [20]Howard Hughes Medical Institute, Chevy Chase, MD, USA. [30]These authors contributed equally: Manik Kuchroo, Jessie Huang, Patrick Wong. [31]These authors jointly supervised this work: Julie G. Hussin, Guy Wolf, Akiko Iwasaki, Smita Krishnaswamy. *A list of members and their affiliations appears at the end of the paper. ✉e-mail: smita.krishnaswamy@yale.edu

iterations in this dynamic condensation process using potential of heat-diffusion for affinity-based trajectory embedding (PHATE), a manifold affinity-preserving dimensionality reduction method, creates Multiscale PHATE embeddings, whereas evaluating connected cells across granularities creates Multiscale PHATE clusters. Furthermore through efficient scalable implementation, we show that we are able to perform visualization and clustering of large-scale data substantially faster than single-scale visualization techniques like *t*-SNE, UMAP or PHATE[10]. Implementing these multigranular and visualization approaches in such a scalable manner, we have created a tool capable of visualizing, clustering and ultimately deriving meaning from rich single-cell datasets.

We showcase our method using 251 blood samples from 168 patients infected with SARS-CoV-2 (ref. [11]) and clinical data from 2,135 patients admitted to Yale New Haven Hospital (YNHH). With our unique multigranular approach, we can produce high-level summarizations and detailed cell type-specific analyses of 54 million of cells, tasks that would take weeks to perform using previous methods. When combined with manifold density estimation (MELD)[12], our approach can identify cellular populations associated with patient outcome across resolutions. At coarse resolutions, we identify T cells to be broadly protective, whereas monocytes and granulocytes are pathogenic. At finer resolution, we identify $CD16^{hi}CD66b^{-}$ neutrophil, $CD14{-}CD16^{hi}HLA\text{-}DR^{lo}$ monocytes, and interferon-γ (IFN-γ)$^{+}$ granzyme B^{+} T helper type 17 (Th17) cells to be associated with patient mortality. While coarse grain analysis reveals that a cell type (e.g., T cells) may be broadly protective, fine-grain analysis reveals that cellular subsets can be pathogenic, highlighting the need for a multiresolution approach. Next, we show that these Multiscale PHATE-derived cellular groupings can be used to predict outcome better than immunologist-curated populations and groupings produced by other graph-based clustering approaches. Finally, to display the generalizability of our approach across data types, we created a multiscale distillation of patients admitted to YNHH. Built from 18 laboratory, clinical and demographic variables, Multiscale PHATE was used to perform multiresolution analysis of patient clinical states and effectively identified lab variables and cellular populations associated with outcomes.

Results

Multiscale PHATE algorithm. Multiscale PHATE combines a data coarse-graining method called diffusion condensation[9] with a manifold-preserving dimensionality reduction method called PHATE[10] to produce multigranular visualizations and clusters of high-dimensional biological data. The Multiscale PHATE algorithm (Methods Alg. 1) can be broken down into four conceptual steps (Fig. 1a):

1. compute a manifold-intrinsic, diffusion potential representation that learns the nonlinear biological manifold as done in PHATE (Methods and Fig. 1a-i);
2. coarse grain this diffusion potential using a fast diffusion condensation process (Methods and Fig. 1a-ii);
3. select meaningful resolutions for downstream analysis with a gradient-based approach (Fig. 1a-iii);
4. visualize condensed diffusion potential coordinates at selected scales via metric multidimensional scaling (MMDS) and analyze coarser-grain resolutions to obtain multiscale clusters (Fig. 1a-iv).

Multiscale PHATE starts by creating a diffusion potential representation **U** of the original data as done by Moon et al[10] and summarized in Methods. Precisely, first, a distance matrix **D** is calculated between all cells based on their ambient measurements. Distance matrix **D** is converted into affinity matrix **K** using an adaptive-bandwidth Gaussian kernel function so that similarity between two cells decreases exponentially with their distance. Next, **K** is row normalized to obtain the diffusion operator **P**, representing the probability distribution of transitioning from one cell to another in a single step. This diffusion operator **P** is raised to t_D, the PHATE optimal diffusion timescale as computed by von Neumann entropy, to simulate a t_D-step random walk over the data graph. Finally, by taking logarithm of $\mathbf{P}^{t_D}$, we calculate the diffusion potential **U** of the data. Previous work has shown that this internal representation computed in PHATE effectively learns the nonlinear geometry of complex biological datasets and can be rapidly visualized in two or three dimensions using MMDS. Multiscale PHATE uses this diffusion potential representation as the substrate for our diffusion condensation process. As done for our diffusion potential calculation, diffusion condensation computes a diffusion operator $\mathbf{P}_t$ at each iteration using a fixed-bandwidth Gaussian kernel function from the location of cells in diffusion potential space. The use of a fixed bandwidth gives a measure of locality in computing cell–cell affinities. This diffusion operator $\mathbf{P}_t$ is applied to the diffusion potential $\mathbf{U}_t$, acting as a diffusion filter, effectively replacing the coordinates of a point with the weighted average of its diffusion neighbors. When the distance between two cells falls below a distance threshold, cells are merged together, denoting them as belonging to the same cluster going forward. This process is then repeated iteratively until all cells have collapsed to a single cluster.

By conducting this denoising over the diffusion potential, Multiscale PHATE tackles two shortcomings of the original diffusion condensation. Diffusion condensation in its original form is not effective at learning or visualizing the nonlinear geometry of biological datasets and is prone to condensing points off the data manifold (Extended Data Fig. 1a). By first learning the nonlinear data manifold through a diffusion potential calculation and feeding this into diffusion condensation, we not only effectively learn the nonlinear geometry of complex datasets (Extended Data Fig. 1a) but also rapidly visualize and learn clusters at resolutions of interest (Fig. 1a-iv).

To identify meaningful scales, we applied a gradient-based approach (Methods), which identifies stable resolutions of the condensation process for downstream analysis. Visualization of any of these resolutions is achieved by computing a potential distance matrix $\mathbf{D}_{\mathbf{U}_t}$ using distance between pairs of rows in $\mathbf{U}_t$. Finally, Multiscale PHATE visualization is obtained by performing MMDS to preserve the distances within $\mathbf{D}_{\mathbf{U}_t}$ in two or three dimensions and ready for visualization. Thus, in Multiscale PHATE, we are able to not only compute a coherent data topology along the data manifold but also quickly visualize an intermediate layer of the condensation process (Extended Data Fig. 1a). Using a stochastic block model, where clusters are known, we show that diffusion condensation initialized with diffusion potential outperforms diffusion condensation on the ambient measurement space as increasing amounts of noise are added to the model (Extended Data Fig. 1b).

Further detail on Multiscale PHATE's generalizability (Extended Data Fig. 2), scalability (Extended Data Fig. 1d) and reproducibility (Extended Data Fig. 1e) can be found in Methods. Finally, additional details on the Multiscale PHATE, how it integrates with other analysis techniques (Fig. 1d and Extended Data Fig. 1c), how the method can be leveraged to create a patient manifold and the algorithm's improved ability to identify pathogenic populations (Extended Data Fig. 3) can be found in Methods.

Comparison of Multiscale PHATE with other methods. Because Multiscale PHATE is a multigranular clustering and visualization tool, we evaluated it against a combination of other visualization and coarse-graining tools using a variety of metrics. To determine the necessity of diffusion condensation to learn data organization, we compared Multiscale PHATE with other clustering methods, including Louvain, Leiden and 0-dimension persistent homology

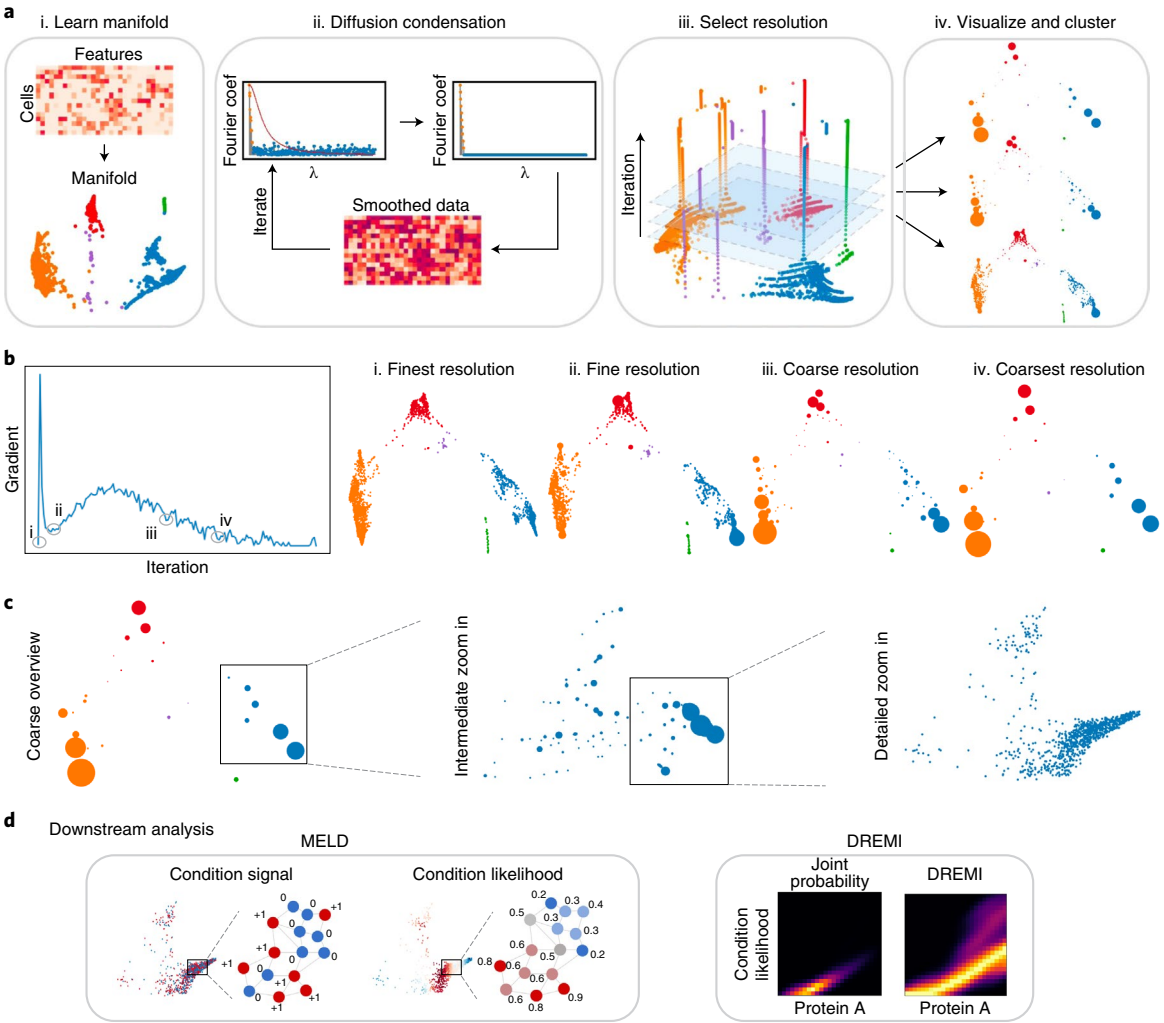


Fig. 1 | Overview of the Multiscale PHATE algorithm. a, Multiscale PHATE process involves four successive steps. The first step (i) learns the manifold geometry via diffusion potential calculation. The second step (ii) iteratively coarse grains the manifold construction through a fast diffusion condensation 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). coef, coefficient. **b**, Gradient analysis identifies a range of scales for visualization by computing shifts in data density from one iteration of the diffusion condensation process to the next. **c**, Multiscale PHATE allows for high-level summarizations of data and zoom ins of data subsets for additional detail. **d**. Multiscale PHATE abstractions of data are amenable to downstream analyses with algorithms like MELD (ref. [12]) and DREMI (ref. [36]).

(single-linkage clustering), using an adjusted Rand index (ARI) and F1 scores as measures of clustering accuracy. Then, with the same data abstraction by each clustering method, we compared our choice of visualization method, PHATE, with UMAP and *t*-SNE. To quantify the visualization by Multiscale PHATE and other comparison combinations, we computed denoised manifold affinity preservation (DeMAP) scores[10] on the embeddings.

Multiscale PHATE embeddings preserved local and global distances. In our comparisons, we performed two different ablation studies to determine the necessity of both the diffusion condensation approach to learn data topology (Fig. 2b) as well as PHATE to learn and visualize manifold geometry (Fig. 2c). In each study, we repeated comparisons on a variety of datasets that have different geometries, such as 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 (Fig. 2a) and running all comparisons, Multiscale PHATE performed better than other methods across nearly all ranges of biological noise (Fig. 2b,c). In particular, Multiscale PHATE had distinct advantages in visualizing data with a high degree of noise (Fig. 2a–c and Extended Data Fig. 4). Although some other methods, such as Homology-UMAP, appear to produce good visualizations, they receive lower DeMAP scores than Multiscale PHATE, suggesting poorer quality. Finally, in our second ablation study (Fig. 2c), it appears that PHATE is the most effective visualization methodology when embedding multiscale clusters generated by the same coarse-graining method. We repeated our comparisons on 1.7 million cells from FlowCap I normal donor (ND) dataset[13], adding increasing amounts of Gaussian noise to simulate variation and increasing degree of undersampling to simulate dropout. Across our comparisons, Multiscale PHATE similarly performed as well or better than other visualization modalities, especially as noise increased within the dataset (Extended Data Fig. 4c,d).

Multiscale clusters accurately captured established groupings of data. To quantify the clustering accuracy of Multiscale PHATE, we benchmarked our approach's ability to predict ground truth clusters on two different types of synthetic data and two different

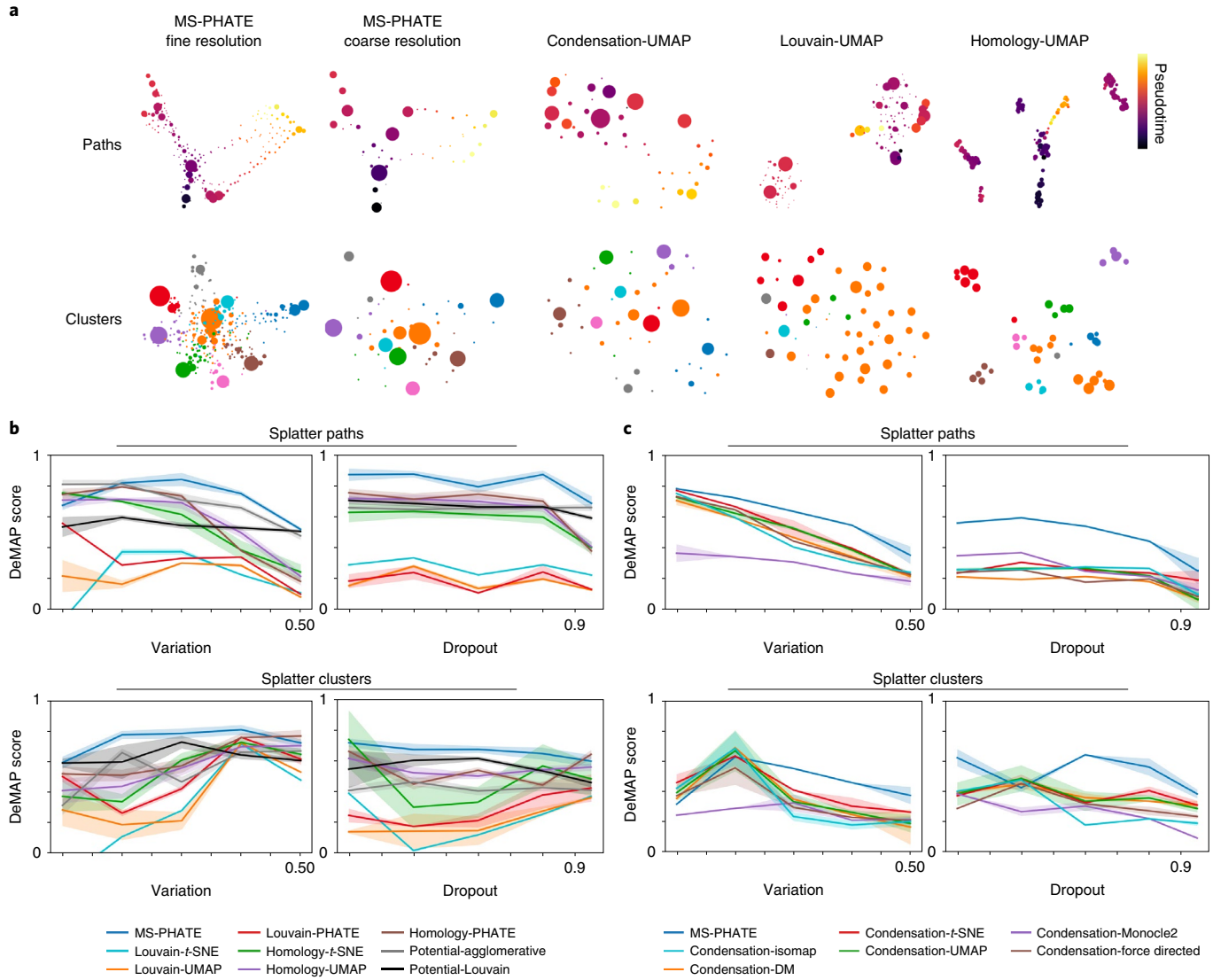


Fig. 2 | Comparison of Multiscale PHATE with other dimensionality reduction tools. a, Visual comparison of Multiscale PHATE (MS-PHATE) with other multiscale dimensionality reduction tools on synthetic single-cell data[14] with either path or cluster structure. In Multiscale PHATE embeddings, each point represents a group of cells that are considered close enough to merge and the size of a dot is proportional to number of cells in that group. Remaining visualizations from multiscale dimensionality reduction tools shown in Extended Data Fig. 4. **b**, Quantitative study comparing embeddings produced by Multiscale PHATE and dimensionality reduction strategies that used either community-based or topologically based abstractions of data. Comparisons were evaluated using DeMAP with increasing levels of two different types of biological noise, dropout and variation, as well as on data with different structures, paths and clusters. Shading represents one standard deviation around the mean DeMAP score for each comparison. **c**, Quantitative study comparing embeddings produced by Multiscale PHATE and alternative dimensionality reduction strategies that visualize condensation-based abstractions of data. Comparisons were run and represented as described in **b**.

types of biological data. First, we simulated noisy synthetic data where ground truth clusters are known, as done previously for visualization comparisons[14]. Then, we computed cluster labels with Multiscale PHATE, Louvain[15], Leiden[16] and single-linkage hierarchical clustering[17] on datasets with varying degrees and types of noise. Across noise levels, Multiscale PHATE outperformed hierarchical, Louvain and Leiden clusterings at the most relevant levels of noise across 10 randomly initialized datasets (Extended Data Fig. 5a). Next, we simulated two- and three-layer hierarchical stochastic block models (Extended Data Fig. 5b). 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. 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, representing a complex form of noise that creates nonlinear changes that would be difficult for many algorithms to deconvolve. Across 10 replicates in three-layer and two-layer models, Multiscale PHATE performed better than Louvain, Leiden and single-linkage hierarchical clustering in 35 of the 42 comparison conditions (Extended Data Fig. 5c,d).

Finally, we benchmarked Multiscale PHATE's performance across granularities on flow cytometry data where cell-type labels have already been established through conventional gating analysis. Across both fine- and coarse-grain cellular clusters, Multiscale PHATE computed clusters that more faithfully represented the underlying known biological cell types (Extended Data Fig. 6a). We next tried to determine whether Multiscale PHATE better captured known populations across a range of computed resolutions.

We computed ARI between known cluster labels and all computed resolutions (less than 100 clusters) of Multiscale PHATE, FlowSOM, Leiden and Louvain. Across all resolutions and both sets of cluster labels, Multiscale PHATE outperformed other models (Extended Data Fig. 6c). Finally, we tried to determine how increasing amounts of noise in real biological data could affect clustering ability. To perform this analysis, we analyzed FlowCAP I ND dataset and added increasing amounts of variation or dropout, computing clusters with all our methods at each noise level. As an increasing amount of noise was added to the data, Multiscale PHATE outperformed other clustering modalities (Extended Data Fig. 6d).

Multiscale PHATE analysis of 251 blood samples from patients with SARS-CoV-2. A total of 168 patients with moderate to severe COVID-19 (ref. [18]) 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 time points to characterize patient cellular responses across the spectrum of disease. In total, the composition of peripheral blood mononuclear cells (PBMCs) was measured by flow cytometry on 251 samples. Finally, clinical data were extracted from the electronic health record corresponding to each biosample time point to allow for clinical correlation of findings (Methods). In this analysis, we define a poor or adverse outcome as a patient who died of infection and a good outcomes as a patient who survived. Rigorous and robust analysis of over 54 million cells characterized across four different sets of flow marker panels is not possible through current single-cell computational techniques. Thus, we applied Multiscale PHATE to identify subsets of PBMCs associated with mortality and survival.

Key cellular subsets were enriched in patients who died of infection. To explore the role of individual PBMC types in disease pathogenesis, we examined 22 million cells measured on a myeloid-centric flow cytometry panel containing samples from 210 patients with COVID-19 across scales with Multiscale PHATE. Using cell type-specific marker staining, we characterized Multiscale PHATE clusters (Fig. 3a). We computed the mortality likelihood score for each patient using MELD with the mortality outcome and identified cellular states enriched in patients who died from infection (darker red) or patients who survived (darker blue) (Fig. 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 ($CD14^{+}$), whereas the population most enriched in survival was T cells ($CD3^{+}$) (Fig. 3c). Although these broad cell types may be associated with disease outcome, cellular subsets likely may be driving some or all of these cell-type effects. We zoomed in on these broad cell types across a number of flow cytometry panels to identify cellular subtypes potentially responsible for pathogenic or protective effects.

$CD14^{-}CD16^{hi}HLA\text{-}DR^{lo}$ monocytes associated with mortality. To identify monocyte subsets implicated in disease, we zoomed into the monocyte population and identified major subtypes based on the expression of markers CD16 and CD14 (Extended Data Fig. 7a). The combination of these markers allowed us to distinguish between $CD14^{+}CD16^{-}$ monocytes, $CD14^{+}CD16^{int}$ monocytes and $CD14^{-}CD16^{hi}$ monocytes. We identified that $CD14^{-}CD16^{hi}$ monocytes were the most strongly enriched in severe infection, followed by $CD14^{+}CD16^{int}$ monocytes (Extended Data Fig. 7b). 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[8,19,20]. Furthermore, across all monocytes, CD16 was positively correlated with mortality, whereas CD14 and HLA-DR were correlated with survival, identifying a distinct $CD14^{-}CD16^{hi}HLA\text{-}DR^{lo}$ population of monocytes enriched in mortality. The loss of HLA-DR on monocytes has been previously observed in patients with COVID-19 and sepsis, potentially via an increase in circulating interleukin-10 (IL-10) (ref. [21]).

Circulating, resting neutrophils associated with mortality. Using Multiscale PHATE, 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) (Fig. 3d). After mapping our mortality scores onto this granulocyte population, we found that $CD16^{hi}$ neutrophils were enriched in patients who died of infection. 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 although 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 who died of COVID-19 (Fig. 3e). Based on the PBMC isolation protocol used (Methods), the neutrophils obtained were by definition low-density neutrophils, containing both the mature and immature subsets. Considering the sensitivity of CD16 expression, the $CD16^{hi}$ neutrophils in our cohort were most likely indicative of a mature population that has not responded to an activating stimulus[22]. Neutrophils from patients with worse disease also expressed less CD66b; in contrast, an increase in surface expression of CD66b occurs following degranulation[23]. Although granulocytes are broadly associated with negative outcomes, Multiscale PHATE reveals that there is actually a subpopulation of circulating resting neutrophils, defined by a combination of high complexity, high CD16 expression and low CD66b expression, that may drive a majority of this pathogenic effect in patients.

Plasmablast populations associated with mortality. In our broad PBMC analysis, B cells were among the most enriched populations in severe outcomes (Fig. 3c). 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 that would otherwise be difficult to detect using traditional two-dimensional gating, a popular method used for flow cytometry analysis (Extended Data Fig. 7c). After identifying these major cell types, we computed mortality likelihood scores to identify B cell subtypes implicated in mortality. The most enriched cell type in patients with adverse outcomes was a subset of the antibody-secreting population defined by $CD86^{lo}HLADR^{-}/CXCR3^{+}$, 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 $CD86^{+}$ (Extended Data Fig. 7d). Despite the protective roles of circulating antibodies, these results are consistent with earlier findings, which discuss potentially pathogenic B cells during COVID-19 infection[24].

Fine-grained analysis identified pathogenic Th17 cells. Although T cells collectively were enriched in patients who recovered from infection (Fig. 3c), there are diverse subsets of T cells that have been implicated in severe disease pathogenesis. To identify functional T cell subsets enriched in patients who died of 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 (Fig. 4a).

Using Multiscale PHATE's zoom and cluster capabilities, we were able to visualize $CD4^{+}$ T cells and subdivide these cells into functional subsets using the functional markers IFN-γ, IL-17 and IL-4 (Fig. 4b). In our dataset, we identified two different subsets of $CD4^{+}$ IL-17-producing T cells classically known as Th17 cells,

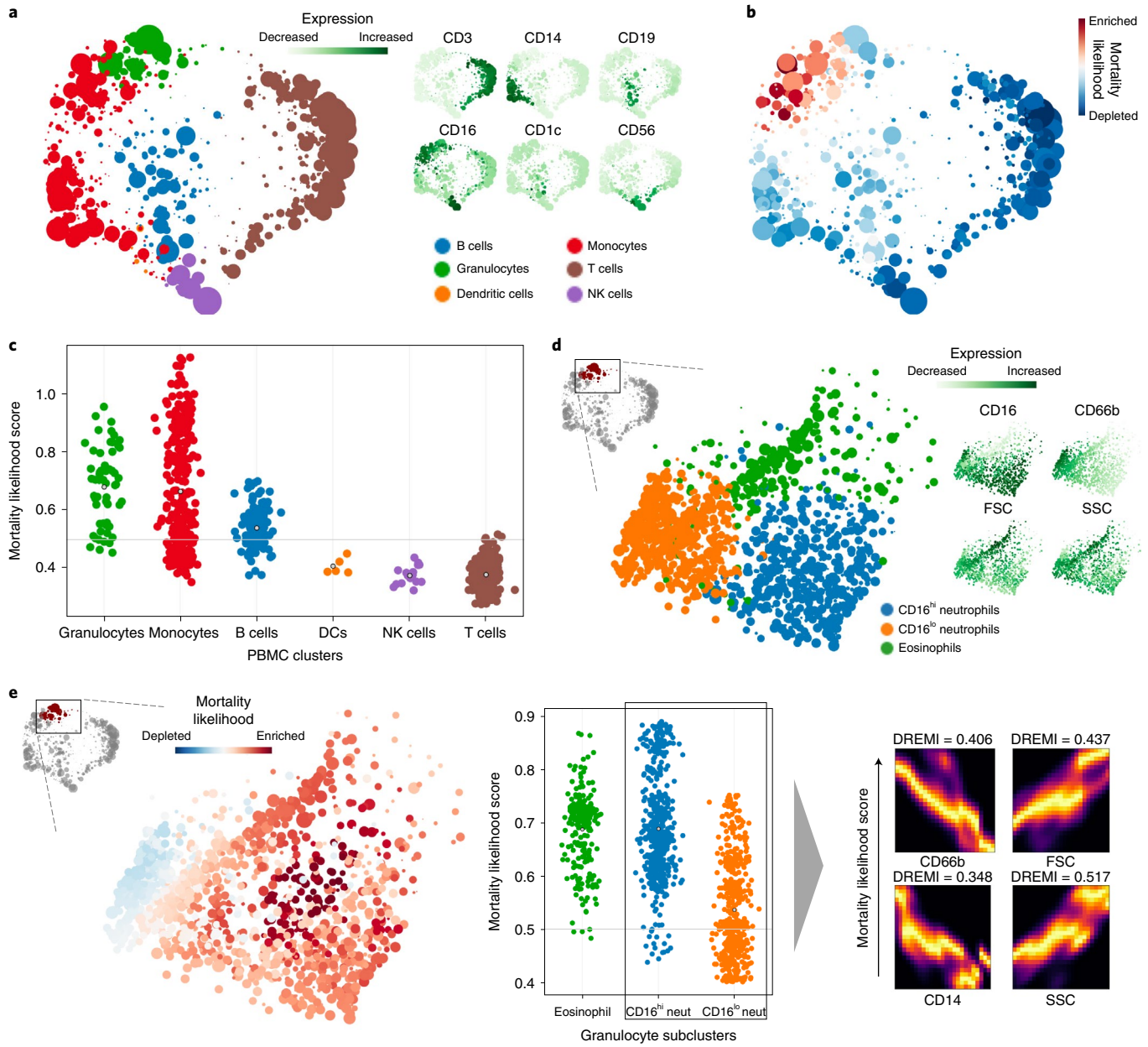


Fig. 3 | The CD16^{hi}CD66b^{lo} neutrophil subset was enriched in patients who died of COVID-19. a, Multiscale PHATE visualization of PBMCs identifies all major cell types based on cell type-specific markers. Colors denote cell type and size of a dot is proportional to number of cells represented. **b**, Visualization of mortality likelihood score computed by MELD on coarse-grain Multiscale PHATE visualization of PBMCs as visualized in **a**. **c**, Visualization of mortality likelihood score computed by MELD organized by cell type revealed enrichment of granulocytes, monocytes and B cells in patients who died of COVID-19. Each dot represents a grouping of cells at the resolution visualized in **a**. **d**, Zoom in of granulocyte population identified subsets of neutrophils and eosinophils based on expression of known markers. **e**, Visualization of mortality likelihood score in granulocyte population identified 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.

one coproducing granzyme B and IFN-γ and one staining negative for both markers. To identify cell types enriched in mortality, we computed a mortality likelihood score. By organizing our scores by Th cell subset, it became clear that the Th17 cell subset coproducing IFN-γ$^+$ granzyme B$^+$ cells was enriched in patients who died of infection. Furthermore, granzyme B and IFN-γ were positively associated with mortality likelihood on DREMI analysis across all CD4$^+$ T cell subsets (Fig. 4c). Although Th17 cells can play protective roles[25], IFN-γ$^+$ granzyme B$^+$ Th17 cells are associated with tissue damage, as observed in models of murine autoimmune encephalomyelitis[26] and neutrophil expansion via IL-17. With COVID-19, this latter mechanism may be relevant given the harmful contribution of and neutrophil extracellular traps during disease[27]. Patients with adverse outcomes in this cohort demonstrated an enrichment in IFN-γ$^+$ granzyme B$^+$ Th17 cells, as well as CD16$^+$ neutrophils. We posit that IFN-γ$^+$ granzyme B$^+$ Th17 cells in our cohort may precipitate these pathogenic effects via IL-17 secretion and subsequent induction of IL-8 from airway epithelial cells or granulocyte colony-stimulating factor from microvascular pericytes[28,29].

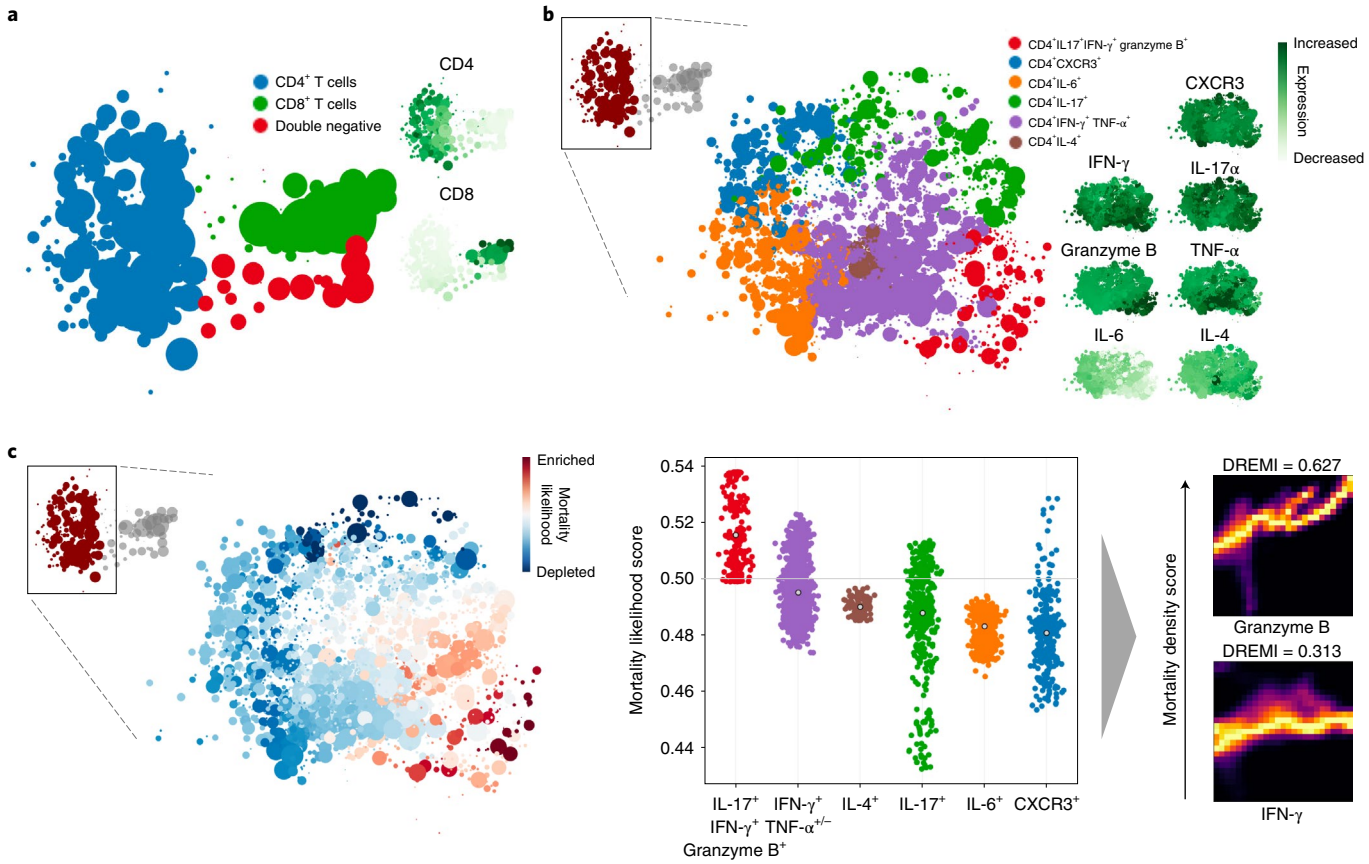


Fig. 4 | Multiscale PHATE identified Th17 cell subsets enriched in patients who died of COVID-19. a, Multiscale PHATE visualization of a T cell-focused cytokine panel identified broad T cell subtypes. Each point is a subgroup of cells, and the size is proportional to the number of cells in the group. **b**, Zoom in of CD4+ Th cells identified subsets based on expression of functional markers. **c**, Visualization of mortality likelihood score computed by MELD identified IFN-γ+ granzyme B+ Th17 cell enrichment in patients with poor outcomes. Key associations between markers and mortality likelihood scores were computed by DREMI and visualized with DREVI. DC, dendritic cell; neut, neturophil; NK, natural killer.

Hyperactivated CD8+ TEMRA cells associated with mortality. Although Multiscale PHATE determined that T cells were broadly protective, we identified a subset of CD4+ T cells that were shown to be pathogenic at finer resolution. Though CD8+ T lymphocytes play a critical role in the clearance of virus during acute illness through the secretion of granzyme B (refs. [30,31]), we tried to determine the differing states present in CD8+ T cells and their role in disease pathogenesis.

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 granzyme B, one producing IFN-γ and one producing tumor necrosis factor α (Extended Data Fig. 8a). After mapping mortality likelihood scores onto the CD8+ subpopulation, it became clear that the granzyme B+ population was most enriched in mortality, as granzyme B expression in CD8+ T cells was highly associated with mortality (Extended Data Fig. 8b). These findings are consistent with a previous study of patients with SARS-CoV-2 that observed an association between CD8+ T cell-derived granzyme B and increased disease severity[32]. To gain additional insight into which discrete subset of CD8+ T cells is the source of granzyme B, 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+, CD8+ and double-positive T cell subsets (Extended Data Fig, 8c); we zoomed into the CD8+ subset and identified a range of activation states based on the expression of key markers (Extended Data Fig. 8d). After computing the MELD mortality likelihood score, we identified that the T Effector Memory re-expressing CD45RA (TEMRA) population displayed the most enrichment in severe infection. Furthermore, across all CD8+ T cells, the activation state markers PD1, TIM3, HLA-DR and CD45RA were also positively correlated with mortality on DREMI analysis (Extended Data Fig. 8e). In agreement with another study of patients with SARS-CoV-2 (ref. [32]), we found a hyperactivated CD8+ T cell response in the form of CD8+CD45RA+TIM3+HLA-DR+PD1+ TEMRA cells likely expressing granzyme B that were correlated with disease lethality.

Patient manifold revealed potential mechanisms of disease. Here, we showed 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. Because Multiscale PHATE creates cellular groupings at multiple granularities, we can derive a rich summarization of patients across scales. Furthermore, it can be useful to use patient data to predict outcome.

We created patient embedding using cluster proportions from several levels of the condensation topology of the myeloid-focused flow cytometry using our patient manifold approach (Fig. 5a and Methods). The resultant embedding demonstrated that patients (or

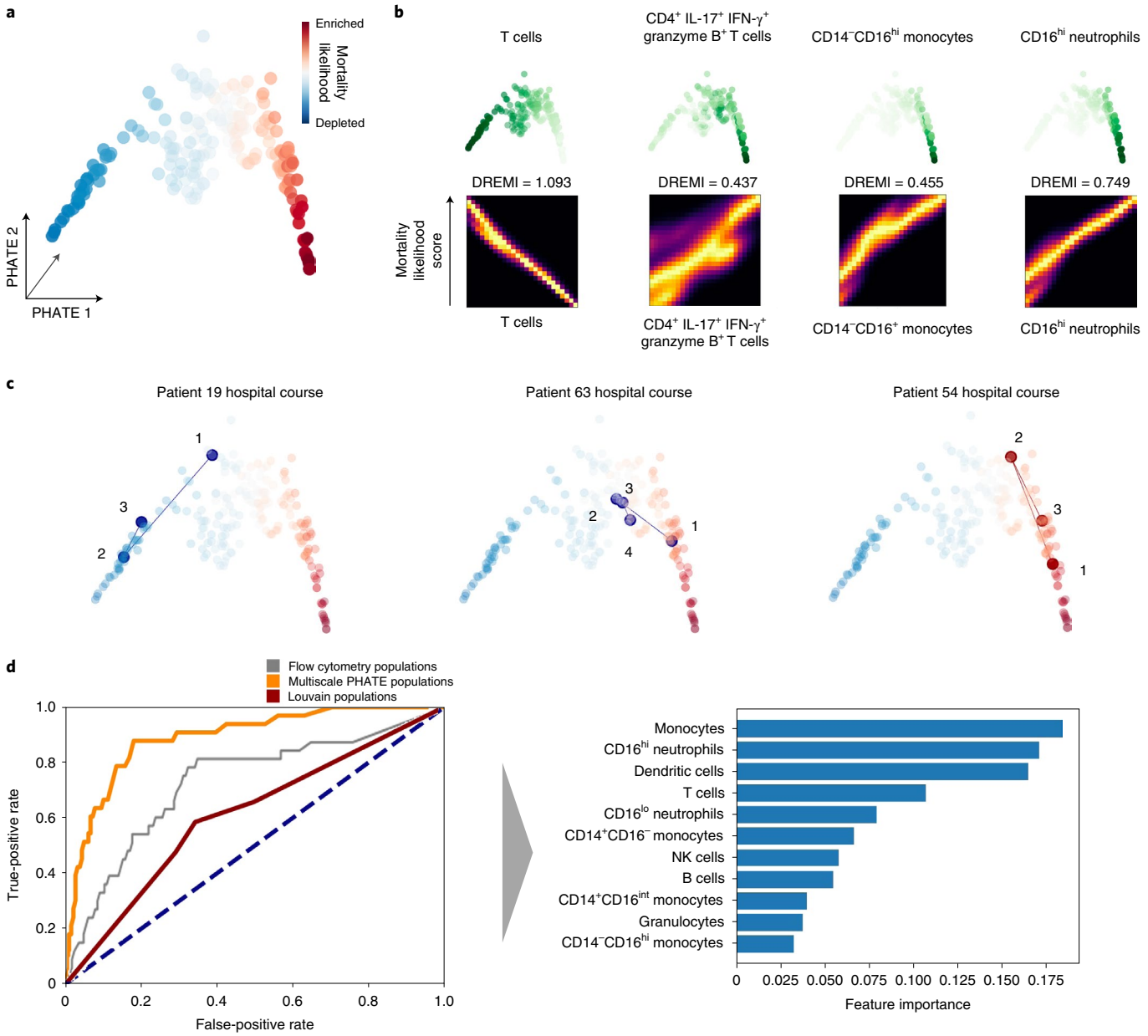


Fig. 5 | Patient manifold corroborated cellular states associated with disease pathogenesis. a, Visualization of patient manifold via PHATE and mortality likelihood score based on patient outcomes computed via MELD. Each point in the PHATE plot represents a patient time point. **b**, Visualization of key cell population enrichment trends over the manifold, with associations computed by DREMI and visualized with DREVI. A darker color in the PHATE plot indicates higher enrichment of the cell type. **c**, Tracing the hospital courses of three patients over the patient manifold. Patients 19 and 63 were discharged, whereas patient 54 died. **d**, Comparing the predictability of patient mortality using random forest classifier on Multiscale PHATE-identified populations, flow cytometry-identified populations and Louvain populations. Accuracy was derived from fivefold cross-validation. The most predictive Multiscale PHATE clusters were ranked using feature-importance analysis.

patient time points) lie on a continuum or manifold themselves. When the patient embedding is colored by the MELD mortality likelihood, we saw that the dominant progression in the data was indeed clinical outcome. We compared our patient manifold construction against a patient manifold constructed from a single resolution of Louvain clustering and conventional flow cytometry gates (Extended Data Fig. 9c). As done in our multiscale approach, we computed feature descriptors of cluster proportions, this time using Louvain partitions and flow cytometry gates as the cellular groupings. Unlike the Multiscale PHATE patient manifold, single-resolution Louvain and flow cytometry patient manifolds representing patients who died of COVID-19 appeared in all regions of the embedding, indicating that this manifold was substantially less meaningful at capturing patient states and outcomes.

To associate previously identified cellular populations with outcome, we computed DREMI between these population proportions and mortality likelihood score. We identified that although T cells were negatively correlated with mortality overall, CD4$^+$ IFN-γ$^+$ granzyme B$^+$ Th17 cells, CD16^{hi} neutrophils and CD14$^-$CD16^{hi} monocytes were strongly positively associated with mortality (Fig. 5b). These findings indicate that a precipitous decline in T cells correlates with mortality, whereas subsets of neutrophils, monocytes and Th17 cells are increased in patients with adverse outcomes. Finally, we traced clinical states of three patients (19, 63 and 54)

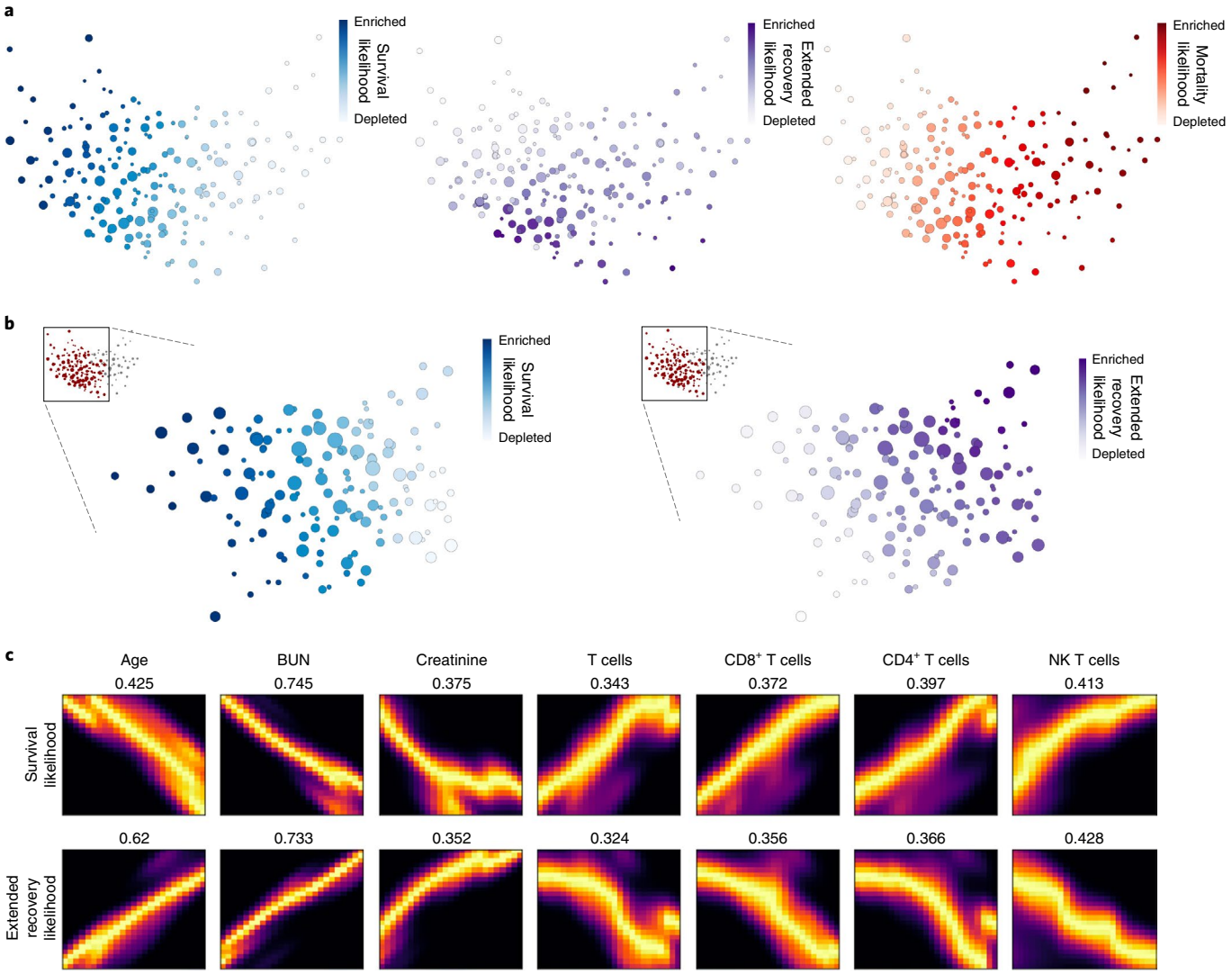


Fig. 6 | Multiscale manifold of patient clinical features identified cell types associated with an extended COVID-19 recovery phase. a, Visualization of a Multiscale PHATE clinical manifold constructed on patient clinical features. Embedding is colored by likelihood scores based on patient outcomes computed via MELD. **b**, Zoom in on the transition point between a high extended recovery likelihood score and a high survival likelihood score. **c**, Patient clinical features and flow cytometry-identified cell populations associated with patient outcomes using DREMI and visualized with DREVI.

across the patient manifold to determine whether our construct accurately recapitulated 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 died of disease, had a tortuous set of clinical states, all of which mapped within the high-mortality region (Fig. 5c). To identify clinical variables associated with mortality, we mapped these patient features onto the manifold, identifying that patients who were older, male, received ventilatory support and had higher markers of inflammation were more likely to experience poor outcomes (Extended Data Fig. 9a). We subsequently 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, which agrees with previous literature demonstrating sex- and age-dependent immune responses[33–35].

To determine whether Multiscale PHATE-derived subpopulations 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 (Methods). Using these abstracted features, we achieved prediction accuracy of $83.7 \pm 0.6\%$ via fivefold cross-validation, with an accuracy of $74.2 \pm 0.8\%$ for mortality cases and $85.5 \pm 0.7\%$ for survival cases. 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 Multiscale PHATE-based classifier model (Fig. 5d). When performing a similar prediction task using flow cytometry-gated populations and Louvain-computed populations, however, we predicted outcome with a lower accuracies of $73.8 \pm 0.8\%$ and $64.7 \pm 1.1\%$, respectively.

Clinical manifold revealed mechanisms of disease convalescence. 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 laboratory, clinical and demographic data generated from routine clinical care of patients with COVID-19 admitted to YNHH. Using 18 clinical and demographic measurements collected on 2,135 patients admitted to YNHH and 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, 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 (Fig. 6a). To understand how clinical features could inform outcomes, we performed DREMI and DREVI analysis between clinical features and each of our likelihood scores (Extended Data Fig. 10a,b). As anticipated, markers of physiologic instability and organ dysfunction (e.g., decreased systolic blood pressure and increased respiratory rate, blood urea nitrogen, creatinine, aspartate aminotransferase and alanine aminotransferase) and systemic inflammatory markers (e.g., increased ferritin, procalcitonin and C-reactive protein) were associated with higher 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 (Fig. 6a). To understand which cellular populations and clinical features drive the difference between these outcomes, we zoomed into this transition point (Fig. 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 (Fig. 6c).

Discussion

Here, we present a multiscale data exploration technique to visualize, cluster and compare large-scale datasets, filling a key gap in biological data exploration. Multiscale PHATE found groupings of data at varying scales that were predictive of clinical outcome. Biological data naturally contain multigranular 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, because of the constant merges that occur in hierarchical clustering approaches (e.g., Louvain), many levels of resolution are missed, and biologically relevant levels of granularity are not recapitulated. In contrast, Multiscale PHATE offers a fast manifold learning-based technique for uncovering a continuum of resolutions of structure and features by understanding data topology. We show that Multiscale PHATE can be combined with other techniques, such as MELD and mutual information (DREMI), to provide deep and detailed insights into biological processes. With Multiscale PHATE, these tools allow users to find resolutions that naturally capture the salient differences between patients, isolate pathogenic and protective cellular subsets across scales and identify key markers associated with disease. T cells, for instance, have been shown to be protective against poor outcomes, corroborating previous work done with COVID-19. Although this cell type is broadly protective, a multiscale zoom in of $CD4^+$ T cells, in combination with MELD and DREMI analysis, reveals a pathogenic $CD4^+$ IFN-γ^+ granzyme B^+ Th17 cell subpopulation. The multiresolution analysis we performed stresses the need to analyze data at multiple granularities. Although 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. Although we have demonstrated Multiscale PHATE in the context of data from patients with COVID-19, both the technique and the ways in which we have used it to analyze a variety of biomedical data, including scRNA-seq, scATAC-seq, cytometry by time of flight, T cell receptor repertoire sequencing and clinical datasets. 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 analysis.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41587-021-01186-x.

Received: 8 March 2021; Accepted: 10 December 2021;
Published online: 28 February 2022

References

1. Klein, A. M. et al. Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. *Cell* **161**, 1187–1201 (2015).
2. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
3. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
4. van der Maaten, L. & Hinton, G. Visualizing high-dimensional data using t-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
5. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotechnol.* **37**, 38 (2019).
6. Pearson, K. LIII. on lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Phil. Mag.* **2**, 559–572 (1901).
7. Lee, J. S. et al. Immunophenotyping of COVID-19 and influenza highlights the role of type i interferons in development of severe COVID-19. *Sci. Immunol.* **5**, eabd1554 (2020).
8. Liao, M. et al. Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat. Med.* **26**, 842–844 (2020).
9. Brugnone, N. et al. Coarse graining of data via inhomogeneous diffusion condensation. *In* Proc. 2019 *IEEE International Conference on Big Data*, 2624–2633 (IEEE, 2019).
10. Moon, K. R. et al. Visualizing structure and transitions in high-dimensional biological data. *Nat. Biotechnol.* **37**, 1482–1492 (2019).
11. Lucas, C. et al. Longitudinal analyses reveal immunological misfiring in severe COVID-19. *Nature* **584**, 463–469 (2020).
12. Burkhardt, D. B. et al. Quantifying the effect of experimental perturbations at single-cell resolution. *Nat. Biotechnol.* **39**, 619–629 (2021).
13. Aghaeepour, N. et al. Critical assessment of automated flow cytometry data analysis techniques. *Nat. Methods* **10**, 228–238 (2013).
14. Zappia, L. et al. Splatter: simulation of single-cell RNA sequencing data. *Genome Biol.* **18**, 174 (2017).
15. Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech.* **2008**, P10008 (2008).
16. Traag, V. A. et al. From Louvain to Leiden: guaranteeing well-connected communities.*Sci. Rep.* **9**, 5233 (2019).
17. Sibson, R. SLINK: an optimally efficient algorithm for the single-link cluster method. *Comp. J.* **16**, 30–34 (1973).
18. Marshall, J. C. et al. A minimal common outcome measure set for COVID-19 clinical research. *Lancet Infect. Dis.* **20**, e192–e197 (2020).
19. Padgett, L. E. et al. Interplay of monocytes and T lymphocytes in COVID-19 severity. Preprint at *bioRxiv* https://doi.org/10.1101/2020.07.17.209304 (2020).
20. Sánchez-Cerrillo, I. et al. COVID-19 severity associates with pulmonary redistribution of $CD1c^+$ DC and inflammatory transitional and nonclassical monocytes. *J. Clin. Invest.* **130**, 6290–6300 (2020).
21. Laing, A. G. et al. A dynamic COVID-19 immune signature includes associations with poor prognosis. *Nat. Med.* **26**, 1623–1635 (2020).
22. Pillay, J. et al. A subset of neutrophils in human systemic inflammation inhibits T cell responses through mac-1. *J. Clin. Invest.* **122**, 327–336 (2012).
23. Fortunati, E., Kazemier, K. M., Grutters, J. C., Koenderman, L. & Van den Bosch, vanJ. M. M. Human neutrophils switch to an activated phenotype after homing to the lung irrespective of inflammatory disease. *Clin. Exp. Immunol.* **155**, 559–566 (2009).
24. Biasi, S. D. et al. Expansion of plasmablasts and loss of memory B cells in peripheral blood from COVID-19 patients with pneumonia. *Eur. J. Immunol.* **50**, 1283–1294 (2020).
25. Kudva, A. et al. Influenza A inhibits Th17-mediated host defense against bacterial pneumonia in mice. *J. Immunol.* **186**, 1666–1674 (2010).

26. Lee, Y. et al. Induction and molecular signature of pathogenic TH17 cells. *Nature Immunology* **13**, 991–999 (2012).
27. & Zuo, Y. et al. Neutrophil extracellular traps in COVID-19. *JCI Insight* **4**, e138999 (2020).
28. Jones, C. E. & Chan, K. Interleukin-17 stimulates the expression of interleukin-8, growth-related oncogene-α, and granulocyte–colony-stimulating factor by human airway epithelial cells. *Am. J. Respir. Cell Mol. Biol.* **26**, 748–753 (2002).
29. Liu, R. et al. IL-17 Promotes neutrophil-mediated immunity by activating microvascular pericytes and not endothelium. *J Immunol* **197**, 2400–2408 (2016).
30. 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. *J. Virol.* **88**, 11034–11044 (2014).
31. Barber, D. L., Wherry, E. J. & Ahmed, R. Cutting edge: rapid in vivo killing by memory CD8 T cells. *J. Immunol.* **171**, 27–31 (2003).
32. Kang, C. K. et al. Aberrant hyperactivation of cytotoxic T cell as a potential determinant of COVID-19 severity. *Int. J. Infect. Dis.* **97**, 313–321 (2020).
33. 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 Immun.* **10**, 509–516 (2009).
34. Klein, S. L. & Flanagan, K. L. Sex differences in immune responses. *Nat. Rev. Immunol.* **16**, 626–638 (2016).
35. McPadden, J. et al. Clinical characteristics and outcomes for 7,995 patients with SARS-CoV-2 infection. *PLoS One* **16**, e0243291 (2021).
36. Krishnaswamy, S. et al. Conditional density-based analysis of T cell signaling in single-cell data. *Science* **346**, 1250689 (2014).

Yale IMPACT Team

Abeer Obaid[21], Adam Moore[22], Alice Lu-Culligan[3], Allison Nelson[21], Anderson Brito[10], Angela Nunez[21], Anjelica Martin[3], Anne L. Wyllie[9], Annie Watkins[10], Annsea Park[3], Arvind Venkataraman[3], Bertie Geng[21], Chaney Kalinich[10], Chantal B. F. Vogels[9], Christina Harden[10], Codruta Todeasa[21], Cole Jensen[10], Daniel Kim[3], David McDonald[21], Denise Shepard[21], Edward Courchaine[23], Elizabeth B. White[10], Eric Song[3], Erin Silva[21], Eriko Kudo[3], Giuseppe Deluliis[21], Haowei Wang[10], Harold Rahming[21], Hong-Jai Park[21], Irene Matos[21], Isabel M. Ott[9], Jessica Nouws[21], Jordan Valdez[21], Joseph Fauver[10], Joseph Lim[24], Kadi-Ann Rose[21], Kelly Anastasio[25], Kristina Brower[10], Laura Glick[21], Lokesh Sharma[21], Lorenzo Sewanan[21], Lynda Knaggs[21], Maksym Minasyan[21], Maria Batsu[21], Maria Tokuyama[3], M. Cate Muenker[21], Mary Petrone[10], Maxine Kuang[10], Maura Nakahata[21], Melissa Campbell[14], Melissa Linehan[3], Michael H. Askenase[26], Michael Simonov[21], Mikhail Smolgovsky[21], Nathan D. Grubaugh[27], Nicole Sonnert[3], Nida Naushad[21], Pavithra Vijayakumar[21], Peiwen Lu[3], Rebecca Earnest[9], Rick Martinello[11], Roy Herbst[21,27,28], Rupak Datta[1], Ryan Handoko[21], Santos Bermejo[21], Sarah Lapidus[9], Sarah Prophet[21], Sean Bickerton[23], Sofia Velazquez[26], Subhasis Mohanty[10], Tara Alpert[1], Tyler Rice[3], Wade Schulz[29], William Khoury-Hanold[3], Xiaohua Peng[21], Yexin Yang[3], Yiyun Cao[3] and Yvette Strong[21]

[21]Yale University School of Medicine, New Haven, CT, USA. [22]Yale University School of Public Health, New Haven, CT, USA. [23]Department of Biochemistry and Molecular Biology, Yale University School of Medicine, New Haven, CT, USA. [24]Yale Viral Hepatitis Program, Yale University School of Medicine, New Haven, CT, USA. [25]Yale Center for Clinical Investigation, Yale University School of Medicine, New Haven, CT, USA. [26]Department of Neurology, Yale University School of Medicine, New Haven, CT, USA. [27]Yale Cancer Center, Yale New Haven Hospital, New Haven, CT, USA. [28]Smilow Cancer Hospital, Yale New Haven Hospital, New Haven, CT, USA. [29]Center of Biomedical Data Science, Yale University, New Haven, CT, USA.

Methods

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.

Multiscale PHATE algorithm. The Multiscale PHATE algorithm is summarized in Alg. 1 as a full integration of PHATE and diffusion condensation.

Algorithm 1. Multiscale PHATE

Input: Data matrix $\mathbf{X}$, kernel parameter ϵ and merge threshold ζ, gradient parameter ϵ

Output: Multiscale PHATE coordinates at T resolutions $\mathbf{J} = \{\mathbf{J}_1, \mathbf{J}_2, \ldots \mathbf{J}_T\}$, selection of scales for visualization S

$[\mathbf{J}_0, \mathbf{U}_0] \leftarrow$ PHATE($\mathbf{X}$)
for $t \in [0, T]$ **do**
  $\mathbf{D}_t \leftarrow$ compute pairwise distance matrix from $\mathbf{U}_t$
  $\mathbf{K}_t \leftarrow$ kernel affinity($\mathbf{D}_t, \varepsilon_t$)
  $\mathbf{P}_t \leftarrow$ row normalize $\mathbf{K}_t$ to get a Markov transition matrix (diffusion operator)
  $\mathbf{U}_{t+1} \leftarrow \mathbf{P}_t\mathbf{U}_t$
  Merge data points i,j if $||\mathbf{U}_{t+1}(i) - \mathbf{U}_{t+1}(j)||_2 < \zeta$, where $\mathbf{U}_{t+1}(i)$ is the i-th row of $\mathbf{U}_{t+1}$
  $\mathbf{D}_{\mathbf{U}t+1} \leftarrow$ compute pairwise distance matrix from $\mathbf{U}_{t+1}$
  $\mathbf{J}_{t+1} \leftarrow$ *MMDS* ($\mathbf{D}_{\mathbf{U}t+1}$)
  $g_{t+1} \leftarrow$ compute gradient from ($\mathbf{U}_{t+1}, \mathbf{U}_t$)
  $\varepsilon_{t+1} \leftarrow$ update($\varepsilon_t, \mathbf{U}_{t+1}$)
end for
for $i \in [1, T-1]$ **do**
  if g_i is a local minimum **then**
    add i to visualization scale set S
  end if
end for

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 (DMs), which rely on diffusion coordinates derived from spectral decomposition of the heat kernel over (Riemannian) manifolds[37]. 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, a diffusion operator is constructed as a matrix with entries $[\mathbf{P}]_{ij} = p(x_i, x_j)$, where $\{x_1, x_2, \ldots\}$ 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 $p^t(x_i, x_j) := \Pr[x_i \underset{t-steps}{\rightarrow} x_j] = [\mathbf{P}^t]_{ij}$. 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.

Although a DM provides appealing analytic relation between spectral embedding with diffusion coordinates[37–39], it often separates trajectories, pathways or clusters into independent eigenspaces. This, in turn, yields multidimensional representations that cannot be conveniently visualized (e.g., having substantially more than two or three dimensions) and cannot be directly projected into two- or three-dimensional displays that faithfully capture diffusion distances. 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 uses 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

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

and the parameter $-1 \le \gamma \le +1$ 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 by Moon et al[10]. There, as well as in other work[40,41], 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. Therefore, the PHATE method is based on embedding potential distances directly into two- or three-dimensional coordinates via a stress-minimizing optimization procedure provided by 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 the study by Moon et al[10].

Multiscale PHATE uses PHATE not only 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.

Multiresolution analysis of diffusion information geometry. 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 $t \to \infty$, all of these distributions converge to a unique equilibrium stationary distribution, as the process is ergodic. PHATE employs an optimal timescale t_D for visualization, which can be identified automatically by distinguishing between a rapid denoising phase and a slow decay from metastable to equilibrium diffusion states. This alleviates the problem of an overly rapid diffusion of information that prohibits multiresolution representation as discussed elsewhere[42,43]. In this paper, we aim 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 by David and Averbuchin[42] as part of a hierarchical construction of localized diffusion folders using a localized diffusion process, which was further analyzed by Wolf et al[43]. The localized diffusion process 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. Although 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 localized diffusion folders. 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 to truly enable multiscale representation learning via diffusion geometry and by extension the diffusion information geometry used in PHATE.

Topological data analysis naturally creates multiscale structure by combining geometric and topological perspectives into a single framework. Although 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.

Learning data topology with diffusion filters in diffusion condensation. A more recent approach toward multiresolution diffusion-based coarse graining was presented in Brugnone et al[9]. Diffusion condensation relies on replacing the traditional time-homogeneous Markov process typically used in diffusion frameworks[37,10] with an inhomogeneous process, following the theoretical analysis in Marshall et al[44] that established the mathematical viability of diffusion geometry construction of such processes. In diffusion condensation, a diffusion operator $\mathbf{P}$ is calculated at each condensation iteration and applied back to an input dataset to slowly condense points toward local centers of gravity as determined by the points diffusion probability between them. This process reduces all eigenvalues besides 1 and diminishes the importance of eigenvectors associated with high-frequency eigenvalues by repeatedly multiplying by a diffusion operator, akin to applying a convolutional filter to the input data, implemented spectrally via a graph Fourier transform, as explained in the following paragraph.

Because the eigenvectors of $\mathbf{P}$, denoted $\Phi = (\phi_0, \phi_1, \ldots, \phi_n)$, represent frequency harmonics over the graph based on a normalized graph Laplacian and graph Laplacian eigenvectors have been shown to be equivalent to graph frequency harmonics[45], signal loadings onto diffusion eigenvectors create a graph Fourier transform defined as $\hat{f} = \Phi^T f$ for a graph signal f. A graph filter can be defined as a rescaling of the coefficients of the graph Fourier transform of a signal. To apply the graph filter to the data, we can apply the graph Fourier transform, rescale the

Fourier coefficients and invert the Fourier transform back to the original space. Thus, a graph filter can be defined by a diagonal matrix H containing rescaled values applied as $\Phi H \Phi^T f$. However, note that the diagonal matrix of eigenvalues of the diffusion operator Λ can itself serve as a low-pass filter. Because $\mathbf{P}$ is a transition matrix of a Markov chain, it has eigenvalues $\lambda_0, \lambda_1, \ldots, \lambda_n$ such that $1 = \lambda_0 \geq \lambda_1 \geq \lambda_n \geq 0$ and thus high-frequency eigenvalues are of lower magnitudes. To apply this diffusion filter to the data, we simply multiply the diffusion operator by the data matrix $\mathbf{P}X$, with $\mathbf{P}X = \Phi\Lambda\Phi^T\mathbf{Q}X$, where X is the data matrix and $\mathbf{Q}$ is a diagonal matrix whose diagonal elements are the row sum of the affinity matrix $\mathbf{K}$. This diffusion condensation process naturally downscales high-frequency and noisy eigenvectors, taking in the whole dataset as the graph signal.

Unlike previous approaches, the coarse graining used in Brugnone et al[9] does not rely on a clustering and 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 toward each other) and fast merging, with concentrated regions collapsing (e.g., as water drops merge together) to single points, 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, then the alternation between these metastable 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.

Condensation on potential coordinates. The computation of the diffusion condensation process described by Brugone et al[9] only uses the diffusion operator $\mathbf{P}$, which is 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 Fig. 1a). As cellular state spaces can be heavily nonlinear[10,36,46], we required an alternative method of diffusion condensation that ensured that the condensed points remain on the manifold. A straightforward method for achieving this might be DM coordinates. However, the computation of DM coordinates requires eigendecomposition of a diffusion operator, which is known to be slow ($O(n^3)$ complexity). In the current paper, rather than using the original features, we used the potential representation of data points used in PHATE (equation (1)) as the as initial features.

The diffusion potential representation, $\mathbf{U}$, of the data is recovered from the transition probabilities of the powered diffusion operator $\mathbf{P}^{t_D}$[10] with the optimal timescale t_D. For the i-th data point, its t_D-step distribution is the i-th row of $\mathbf{P}^{t_D}$ and its potential representation is the i-th row of $\mathbf{U}$. Intuitively, a smaller potential distance corresponds to higher similarity in that it takes less time to diffuse between the point pair. By taking logarithm of $\mathbf{P}^{t_D}$, we allow faraway data points to inform the local distances and balance local and global geometry of the representation. This is the prominent advantage of using diffusion potential instead of directly using the data distribution $\mathbf{P}^{t_D}$, which is found particularly useful for visualizing biological data[10]. This effectively re-represents points by features that consist of the $\log$ of diffusion probabilities to all other features. 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. Using default settings, diffusion condensation is calculated on potential distance using a fixed-bandwidth Gaussian kernel, where the initial bandwidth is set to 1/10 of Silverman's rule of thumb for kernel bandwidth[47]. The bandwidth is then increased by a ratio of 1.025 every iteration.

Scalable coarse graining with fast diffusion condensation. In order to allow Multiscale PHATE to enable scalable exploration of large datasets, such as high-dimensional biological data, we propose speeding up of the initial condensation iteration in the following ways: (1) speeding up the initial iteration using graph partitioning, (2) fast computation of the diffusion potential via landmarking and (3) merging of data points to increase computational efficiency over iterations.

The complexity of computing a diffusion operator on n points is n^2. To reduce n for initial condensation iterations, we run hierarchical k means 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 k-means 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 by Moon et al[10] and explained below) on the centroid of each of these clusters before starting the coarse-graining process.

Instead of using spectral clustering on the full dataset, we came up with cluster centroids that were treated as 'landmarks'. Transition probabilities were computed between points and landmarks and then used with the diffusion potential of landmarks to recover the diffusion potential of all data points. Moon et al[10] showed that this leads to high-quality approximations of the diffusion operator, which leads to near-identical visualizations with PHATE. In addition, we previously found that this leads to low-error approximations of diffusion operators in general[48]. We used this fast approach to compute a low-error diffusion potential system for our coarse-graining process. By default, diffusion potential is calculated using an alpha decay adaptive-bandwidth kernel, which sets its bandwidth to the fifth farthest neighbor in the graph, as originally done by Moon et al[10].

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 ζ), we merge them into a cluster, as they would then have approximately the same coordinates. Using default settings, the merge threshold is set to the 1% smallest distance between any two points in potential space. 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, including DM, t-SNE, UMAP, Monocle 2 and PHATE (Extended Data Fig. 1d).

Selection of visualization scales via gradient analysis. The iterative coarse graining via diffusion condensation generates hundreds of layers for downstream analysis. We propose to select salient levels of granularities for visualization based on gradient analysis. These salient layers of representation must be stable levels that persists for several iterations. To find such levels, we examine the gradient of points of diffusion potential $\mathbf{U}$ across successive condensation iterations and determine where the overall shift in data density from one iteration to the next is locally minimal (Fig. 1b). More specifically, the gradient matrix after a condensation step t is defined as

$$\mathbf{G}_t = \mathbf{U}_t - \hat{\mathbf{U}}_{t-1},$$

where $\hat{\mathbf{U}}_{t-1}$ is computed from $\mathbf{U}_{t-1}$ by taking the average of any subset of rows that are merged during condensation step t to match the dimensions of $\mathbf{U}_t$. If no merges or shifts in data occurred during step t, $\hat{\mathbf{U}}_{t-1} = \mathbf{U}_{t-1}$. The gradient value is then computed by taking the sum

$$g_t = \sum_{i,j} |\mathbf{G}_t(i,j)|.$$

Generally, the gradient changes smoothly from one iteration to the next as semistable resolutions are reached. We pick scales for visualization by identifying local minima in $\{g_1, g_2 \ldots g_T\}$, as observed in the gradient curve (Fig. 1b). Because Multiscale PHATE can compute PHATE embeddings at all condensation steps, visualization at any granularities identified by gradient analysis is readily available (Fig. 1c).

Distinction 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 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[49] or minimum spanning tree-based clustering[50]) 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 this subspace into further clusters. 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. Although 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. Furthermore, this approach corresponds to a very specific epsilon schedule and kernel format (e.g., determined by the used linkage type).

The condensation process used here, on the other hand, is derived from a continuous process that gradually eliminates local variability in the data using a

more gradually changing epsilon schedule and kernel format, which allows for exploration of a more continuous range of granularities. 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 metastability can be analyzed, as we do for the selection of metastable resolutions (e.g., for visualization) explained below.

To demonstrate the difference between diffusion condensation and agglomerative clustering, we use the Louvain method[15] as a representative example because of 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). Although the forced merges ensure a hierarchy of data agglomerations, they do not provide reliable coarse-grained representations for revealing varied data resolutions. As we show in Extended Data Fig. 5, 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 Fig. 5e).

Comparison of multigranular clusters. 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 ARI. We first generated synthetic single-cell data with ground truth cluster labels using Splatter[14]. We then produce a range of noisy splatter datasets, each with increasing amounts of either dropout or variational noise, and run Multiscale PHATE, Louvain[15], Leiden[16] and single-linkage hierarchical clustering[17] to identify groupings across multiple levels of granularity. For each technique at each noise level, we compute ARI between clusters computed across all granularties and ground truth clusters, saving the highest ARI (Extended Data Fig. 5a).

Next, we generated a hierarchical stochastic block model with different clusters at multiple granularities (Extended Data Fig. 5b). We then used Multiscale PHATE, Louvain[15], Leiden[16] and single-linkage hierarchical clustering[17] 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. Finally, for the flow cytometry data, we used gated populations from three 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 part of three distinct populations. ARI was computed similarly for this dataset, and ground truth labels were compared with all granularties of clusters from each algorithm, with the top score stored for each approach (Extended Data Fig. 6c). Networkx[51] was used to produce Louvain clusters, Leidenalg was used to produce Leiden clusters and agglomerative clusters were produced using sklearn[52].

Comparison of multigranular visualizations. To show that Multiscale PHATE created improved multigranular visualizations when compared to other approaches, we presented examples of visualization for qualitative comparison and performed two ablation studies for quantitative comparison. First, Splatter software was used to simulate ground truth and noisy single-cell data of either group (cluster) or path (trajectory) geometries[14]. We showed Multiscale PHATE visualizations of both fine and coarse resolutions on both splatter paths and clusters data to demonstrate our method's ability to visualize at varied granularity. Both resolutions were gradient salient based on the gradient analysis described in the previous section. A fine resolution was chosen to display 200 points, whereas a coarse resolution was chosen to display about 50 points. We compared this method with UMAP visualization of other multiscale abstraction methods, including diffusion condensation, Louvain and computational homology. The resolution of comparison methods in Fig. 2a were chosen to most closely match Multiscale PHATE fine resolution. It should be noted that Louvain only returns a few resolutions (usually only two or three), whereas Multiscale PHATE generates a much wider range of resolutions. The fine granularity of Louvain was the closest match for Multiscale PHATE fine resolution. As for the homology method, we can explicitly set the resolution to match the Multiscale PHATE fine resolution. The same resolution selection strategy for comparison methods applies to the following quantitative comparisons.

We performed two ablation studies, the first to show the necessity of diffusion condensation to learn data topology and the second to show the necessity of PHATE for visualization. In the first ablation study, different approaches used to build a multiscale abstraction of the noisy synthetic data were computed, including diffusion condensation, Louvain and computational homology, as well as Louvain and homology constructed from diffusion potential. Across all methods that use diffusion potential, diffusion potential coordinates were computed using default settings in PHATE (five nearest neighbors, 40 α, 1 γ). Louvain or homology clusters were then computed using these diffusion potential coordinates as the substrate instead of the raw data values. Finally, these abstractions were visualized with a range of dimension reduction and visualization strategies, including PHATE, *t*-SNE and UMAP. For techniques that use diffusion potential for the calculation of clusters (as done by potential-agglomerative and potential-Louvain), all data points corresponding to each cluster at the specified resolution were merged together to form aggregated points (essentially by averaging their feature values). These aggregated points were then visualized with each dimensionality reduction technique.

The resultant embeddings were compared with Multiscale PHATE using DeMAP (ref. [10]). DeMAP is a metric for assessing visualization quality in terms of its ability to capture the manifold geometry of noisy data[10]. 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. We applied each combination of methods to the splatter cluster and path data with increasing levels of two types of noise, variation and dropout, and we calculated the DeMAP score at selected resolutions. The resolution was selected for Multiscale PHATE via gradient analysis and is the same as the fine resolution shown in Fig. 2a. To get a fair comparison, we identified resolutions for Louvain and homology that matched Multiscale PHATE fine resolution most closely at each noise level, respectively.

In the second ablation study, condensation topology on the noisy synthetic data was computed via diffusion condensation initialized with diffusion potential, and an embedding was created after identifying the gradient salient fine resolution via gradient analysis. In order to create multiscale visualizations with other dimensionality reduction strategies, we first aggregated all data points in the ambient space that belong to a Multiscale PHATE cluster at the gradient salient fine resolution as done previously and applied a range of other visualization approaches, including *t*-SNE, Monocle 2, isomap, UMAP, force directed and DM 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 dropout and a range of noise levels.

For robustness, all processes 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[10].

Additional datasets and noise simulation. FlowCAP I ND dataset contains 10-dimensional data from 30 samples with approximately 60,000 cells per sample and a total of over 1.7 million cells. The clustering task is to detect seven manually gated populations. Further details on the dataset are available from the FlowCAP website (http://flowcap.flowsite.org/).

We created two types of noise on this dataset for our clustering and visualization comparisons: biological variation and dropout. We simulated dropout noise on datasets by subtracting random values sampled from a Gaussian distribution to achieve a global undersampling of the data ranging from 10% to 95%. Variation was simulated by adding Gaussian noise to each dimension, ranging from 10% to 50% of the maximum value in each dimension.

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 used 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 evaluated clusters at multiple levels of the condensation topology.

Practically, we created a manifold of samples by simultaneously evaluating multiple levels of the diffusion condensation topology. At each level $\ell \in \{1, 2, \ldots, L\}$, a number of N_ℓ clusters were identified. We counted the number of cells, $n_{\ell,j,k}$, of the k-th patient that belong to each cluster $C_{\ell,j}$ for every $j \in \{1, 2, \ldots, N_\ell\}$ and calculated the normalized percentage as $r_{\ell,j,k} = \frac{n_{\ell,j,k}}{\sum_j n_{\ell,j,k}}$. We calculated the proportions for all patients at a series of selected levels of the topology and concatenated these to create a rich multiscale vector of features for each patient. These multiscale feature vectors were then used to create an embedding with PHATE (ref. [10]) and denoise patient-specific signals using MAGIC (ref. [46]) using Euclidean distance between samples.

By evaluating cluster proportions across multiple resolutions, we created high-dimensional multiscale feature descriptors for each patient that can then be embedded with PHATE for visualization, MELD for outcome likelihood inference and finally DREMI for association analysis (Fig. 5a,b). The constructed patient manifold accurately recapitulated the clinical states (Fig. 5c,d) and better represented patient states than patient manifolds constructed from Louvain clusters and flow cytometry gates (Extended Data Fig. 9c).

Generalizability, scalability and reproducibility of Multiscale PHATE. Multiscale PHATE is broadly generalizable to a large number of biological data

types, including flow cytometry, scRNA-seq, scATAC-seq and clinical variables, among others (Extended Data Fig. 2). When comparing run times between different techniques, it became clear that Multiscale PHATE was able to rapidly scale to millions of cells, successfully embedding 5 million cells in less than 10 min, whereas the next most scalable technique, Monocle 2, could only embed 500,000 cells in a comparable time frame (Extended Data Fig. 1d). Across all comparisons, the number of features did not alter run time drastically, as the initial step of each of these dimensionality reduction algorithms is feature compression with PCA. Thus, the only major difference in run time was the length to compute PCA compression, which is done via a rapid randomized single value decomposition process. Finally, Multiscale PHATE is highly reproducible. A common issue with UMAP and *t*-SNE, which shift clusters from run to run based on initialization, is addressed by Multiscale PHATE, which can faithfully create the same embedding across multiple runs with different initializations (Extended Data Fig. 1e).

Use of MELD with Multiscale PHATE. MELD is a method proposed by Burkhardt et al[12] 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 and a heat kernel at each point (Fig. 1c). In order to apply MELD to this dataset, we combined the flow cytometry data 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 gets 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.

Because Multiscale PHATE identifies clusters of cells across all levels of granularity, we could sweep across resolutions to identify levels that 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 found that Multiscale PHATE was better able to isolate high- and low-mortality likelihood score regions in one of our flow cytometry panels (Extended Data Fig. 5e). By looking at these informative resolutions, we identified populations of cells that were pertinent to patient outcomes. When identifying these subpopulations in conjunction with cell type-defining markers, we found that we could identify cell types and functional subtypes that were differentially enriched across patient outcomes and may drive disease pathogenesis. The full Multiscale PHATE and MELD integrated pipeline is shown in Extended Data Fig. 1c.

DREMI associations with mortality likelihood score. DREMI (ref. [36]) 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, $X=\{X_1, X_2, \ldots, X_n\}$, and $Y=\{Y_1, Y_2, \ldots, Y_n\}$, 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 $\mathrm{DREMI}(X,Y)=I(X,Y|X)=H(Y|X)-\sum_i H(Y|X_i)$. 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 identified functional marker trends that 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 identified markers and, by extension, activation states that are associated with outcome.

Multiscale PHATE improves on current methods to identify and extract pathogenic populations from large biological manifolds. Multiscale PHATE is able to not only visualize and cluster large biological manifolds but also better identify and extract populations of interest in crowded submanifolds. All dimensionality reduction methods suffer from crowding as a result of squeezing high-dimensional data into low-dimensional axes. In crowded regions, it can be difficult to resolve fine-grained structure and separations. The multiscale approach of Multiscale PHATE alleviates crowding by zooming into crowded regions and revealing finer-grained structure (Extended Data Fig. 3a–d). We showcased the utility of our approach by zooming into a crowded region of our PBMC dataset using both Multiscale PHATE and PHATE (Extended Data Fig. 3a–d). Although zooming in clearly separates differing cell types from one another in our Multiscale PHATE approach, it does not in PHATE. Furthermore, when MELD is used in these crowded submanifolds, extracting pathogenic populations with vertex frequency clustering is problematic due to lack of natural separations of the data[12]. When trying to identify populations of cell enriched in patients who died of infection, we clearly identify a subpopulation of B cells enriched in lethal disease on both PHATE and Multiscale PHATE (Extended Data Fig. 3e,f). However, Multiscale PHATE clustering better isolates this population (Extended Data Fig. 3g,i) and furthermore, because of the hierarchical nature of Multiscale PHATE clusters, produces a gating strategy capable of isolating this population (Extended Data Fig. 3h). Altogether, this analysis reveals that Multiscale PHATE is able to better alleviate crowding problems in high-dimensional data, allowing for the identification of sortable pathogenic populations that cannot be done with current baseline methods.

Patient manifold analysis from Multiscale PHATE features. To identify the differences between individual patient samples, we used Multiscale PHATE to construct a manifold of patients as described above. Similar to the mortality likelihood score computed by MELD in our flow cytometry analysis, we computed a similar mortality likelihood score for our patient manifold by identifying whether each patient sample originated from a patient who had a positive outcome or a negative outcome. 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 denoised our patient sample features using MAGIC (ref. [46]) before running association analysis between features using DREMI (ref. [36]).

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 whether the populations we identified across granularities were predictive of patient outcome. To predict patient outcomes from 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 five sets, we performed fivefold cross-validation in which we iteratively shuffled training sets (four of five) and test sets (one of five).

Preprocessing of patient flow cytometry data. Flow cytometry was performed on PBMCs from each patient over the course of several weeks (the methods are explained in detail below). Because of the extended period of patient sample processing, the settings of the flow cytometry could change subtly day to day, producing differences in the amount of fluorescence measured from sample to sample. Because we wanted the distances between cells to reflect real biology instead of experimental artifacts, the normalization steps that we took aimed to ensure that each cell had equal total fluorescent counts.

The resulting FCS files were preprocessed by applying compensation based on the respective single-color compensation controls, selecting only leukocytes and singlets based on FSC and SSC and selecting only live cells based on a viability dye. Mean fluorescence intensity values for each fluorophore on a per-cell basis were then extracted for downstream analysis. 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 2,700 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 2,700 across were removed. The total fluorescent counts are affected by experimental settings and vary substantially between cells. Therefore, we normalized total fluorescent count to 1,000 per cell so that each cell had equal total counts. We then applied square-root normalization to each entry of the data matrix. The normalization for a data matrix $\mathbf{D}$ with n samples and d features is

$$\mathbf{D}_{norm}(i,j)=\left(1000\times\frac{\mathbf{D}(i,j)}{\sum_{k=1}^{d}\mathbf{D}(i,k)}\right)^{1/2}.$$

Biological and medical methods. In the following sections, we provide details on how patient biological data and clinical information were acquired and processed.

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.

Patients. Patient enrollment, sample acquisition, processing and downstream analysis by flow cytometry were performed as in Lucas et al[11]. One-hundred and sixty-eight patients admitted to YNHH with SARS-CoV-2 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-CoV-2 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 at each longitudinal time point. For all patients, days from symptom onset were estimated as follows: (1) the highest priority was given to explicit onset dates provided by patients; (2) the 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 records 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 enzyme-linked immunosorbent assay. 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.

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 1,200*g*. 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 Scientific, AMQAX1000).

Flow cytometry. In brief, freshly isolated PBMCs were plated at 1–2 × 106 cells per well in a 96-well U-bottom plate. Cells were resuspended in Live/Dead Fixable Aqua (Thermo Fisher Scientific) for 20 min at 4 °C. Following a wash, cells were blocked with Human TruStain FcX (BioLegend) for 10 min at room temperature. Cocktails of desired staining antibodies were added directly to this mixture for 30 min at room temperature. 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. Before analysis, cells were washed and resuspended in 100 µl of 4% paraformaldehyde for 30 min at 4 °C. For intracellular cytokine staining following stimulation, cells were resuspended in 200 µl cRPMI (RPMI-1640 supplemented with 10% FBS, 2 mM L-glutamine, 100 U ml^{-1} penicillin, and 100 µg ml^{-1} streptomycin, 1 mM sodium pyruvate and 50 µM 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. Then, 50 µl 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% paraformaldehyde 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 (Thermo Fisher Scientific). Data were analyzed using FlowJo software v10.6 software (Tree Star).

Acquisition of clinical data for flow cytometry analysis and patient manifold. Longitudinal patient data were extracted from the electronic medical record (Epic) only for the patients who were hospitalized and included in the repository. Time-varying data, specifically vital signs and laboratory studies, were extracted specifically 24 h 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 and 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 value) in the 72-h period and overlaid onto the patient manifolds.

Acquisition of clinical data for clinical manifold. For patients who did not undergo flow cytometry analysis, the time-varying clinical, laboratory and treatment data were extracted for the first 24 h 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.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Raw flow cytometry data and count matrices have been deposited to ImmPort and are available through study number SDY1886.

Code 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).

References

37. Coifman, R. R. & Lafon, S. Diffusion maps. *Appl. Comput. Harmon. Anal.* **21**, 5–30 (2006).
38. Bermanis, A., Wolf, G. & Averbuch, A. Cover-based bounds on the numerical rank of Gaussian kernels. *Appl. Comput. Harmon. Anal.* **36**, 302–315 (2014).
39. Belkin, M. & Niyogi, P. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Comput.* **15**, 1373–1396 (2003).
40. Duque, A. F., Wolf, G. & Moon, K. R. Visualizing high dimensional dynamical processes. In *Proc. 2019 IEEE 29th International Workshop on Machine Learning for Signal Processing* https://doi.org/10.1109/MLSP.2019.8918875 (IEEE, 2019).
41. Gigante, S., Charles, A. S., Krishnaswamy, S. & Mishne, G. Visualizing the phate of neural networks. In *Advances in Neural Information Processing Systems 32* (eds Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E. & Garnett, R.) (Curran Associates, Inc., 2019).
42. David, G. & Averbuch, A. Hierarchical data organization, clustering and denoising via localized diffusion folders. *Appl. Comput. Harmon. Anal.* **33**, 1–23 (2012).
43. Wolf, G., Rotbart, A., David, G. & Averbuch, A. Coarse-grained localized diffusion. *Appl. Comput. Harmon. Anal.* **33**, 388–400 (2012).
44. Marshall, N. F. & Hirn, M. J. Time coupled diffusion maps. *Appl. Comput. Harmon. Anal.* **45**, 709–728 (2018).
45. Shuman, D. I., Narang, S. K., Frossard, P., Ortega, A. & Vandergheynst, P. The emerging field of signal processing on graphs: extending high-dimensional data analysis to networks and other irregular domains. *IEEE Signal Process. Mag.* **30**, 83–98 (2013).
46. van Dijk, D. et al. Recovering gene interactions from single-cell data using data diffusion. *Cell* **174**, 716 – 729 (2018).
47. Silverman, B. *Density Estimation for Statistics & Data Analysis* (Chapman & Hall, 1986).
48. Gigante, S. et al. Compressed diffusion. In *Proc. 13th International Conference on Sampling Theory and Applications* https://doi.org/10.1109/SampTA45681.2019.9030994 (IEEE, 2019).
49. Savaresi, S. M. & Boley, D. L. On the performance of bisecting k-means and pddp. In *Proc. 2001 SIAM International Conference on Data Mining* (eds. Kumar, V. & Grossman, R.) https://doi.org/10.1137/1.9781611972719.5 (SIAM, Philadelphia, PA, 2001).
50. Grygorash, O., Zhou, Y. & Jorgensen, Z. Minimum spanning tree based clustering algorithms. In *Proc. 2006 18th IEEE International Conference on Tools with Artificial Intelligence*, 73–81 (IEEE, 2006).
51. Hagberg, A. A., Schult, D. A. & Swart, P. J. Exploring network structure, dynamics, and function using networkx. In *Proc. 7th Python in Science Conference* (eds. Varoquaux, G., Vaught, T. & Millman, J.) 11 – 15 (SciPy, Pasadena, CA, 2008).
52. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).

Acknowledgements

We thank the Mila COVID-19 task force for fruitful discussions and feedback during the conception, development and application of the methods presented here. This study was supported by the Beatrice Kleinberg Neuwirth Fund; the Sendas Family Fund, Yale School of Public Health; and Department of Internal Medicine at the Yale School of Medicine. This work was partially funded by the Institute for Data Valorisation: IVADO Professor funds (G.W.) and COVID-19 Rapid Response grant CVD19-030 (J.G.H.); the Montreal Heart Institute Foundation (J.G.H.); the Canadian Institute for Advanced Research (Canada CIFAR AI Chair) and the Natural Sciences and Engineering Research Council of Canada (NSERC Discovery grant 03267) (G.W.); Merck Investigator Studies Program 60293 (S.F.); the Chan-Zuckerberg Initiative (grants CZF2019-182702 and CZF2019-002440) (S.K.); the National Institute of Health grants 1F30AI157270-01 (M.K.), R01GM135929 (G.W., M.J.H. and S.K.), R01GM130847 (G.W. and S.K.), R01AI157488 and K23MH118999 (S.F.), 1K23DK125718-01A1 (D.S.) and R01DK113191, P30DK079310 and R01HS027626 (F.P.W.); the National Science Foundation (grant DMS1845856) (M.J.H.) and NSF CARRER grant 2047856 (S.K.); and the Sloan Fellowship FG-2021-15883 (S.K.). The content provided here is solely the responsibility of the authors and does not necessarily represent the official views of the funding agencies.

Author contributions

Conception: M.K., J.H., P.W., J.-C.G., S.K., A.I., G.W., J.H., S.F., C.S.D.C., A.I.K., J.G.H., M.J.H. and P.W.; design of work: M.K., J.H., J.-C.G., D.S., A.T., S.K., A.I., G.W. and 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.-M. and J.F.; analysis of data: M.K., J.H., P.W., J.-C.G., D.B.B., A.T., S.G., A.G., F.P.W. and B.R.; interpretation of data: M.K., J.H., P.W., J.-C.G., S.K. A.I., G.W., J.H., S.F., C.S.D.C., A.I.K., F.P.W., J.G.H. and P.W.; creation of new software: M.K. and J.H.; writing: M.K., J.H.,, S.K., A.I., G.W., S.F., C.S.D.C., A.I.K. and P.W.

Competing interests

A.I. served as a consultant for 4BIO Capital and is a cofounder of RIGImmune and Xanadu Bio. S.K. is on the scientific advisory board of KovaDx and AI Therapeutics. P.F.W. is founder of Efference. The remaining authors declare no competing interests.

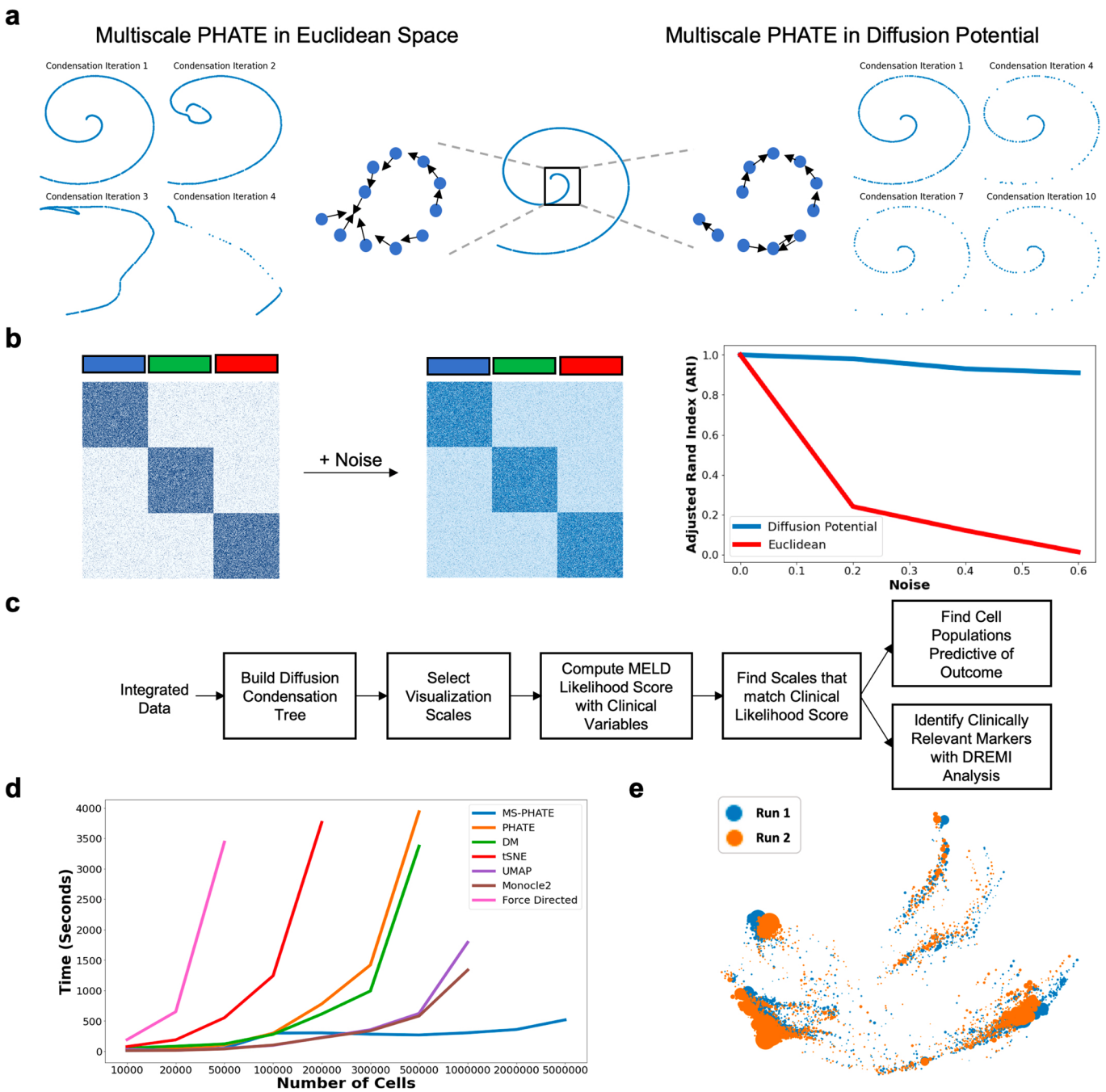
Additional information

Extended data is available for this paper at https://doi.org/10.1038/s41587-021-01186-x.

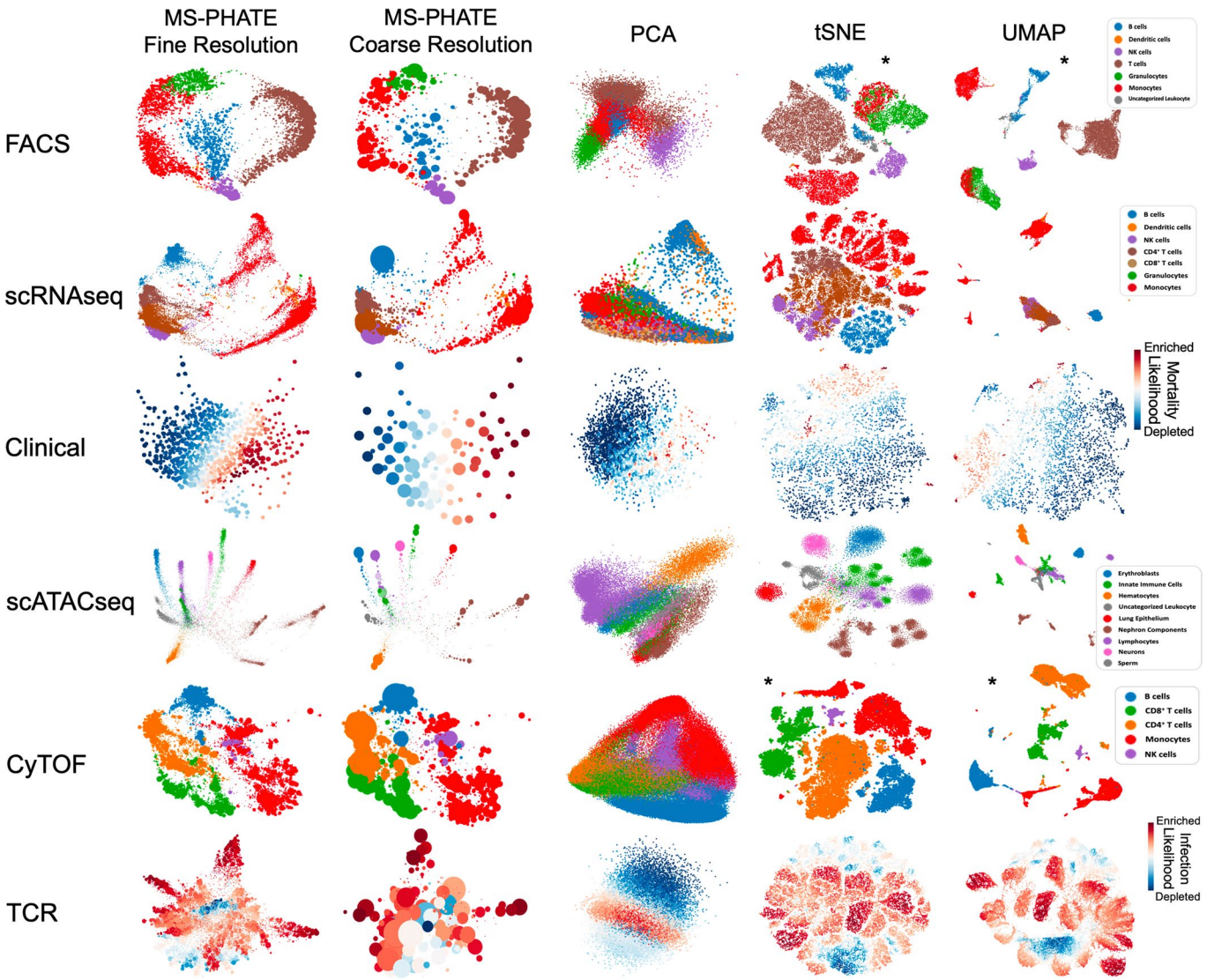
Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-021-01186-x.

Correspondence and requests for materials should be addressed to Smita Krishnaswamy.

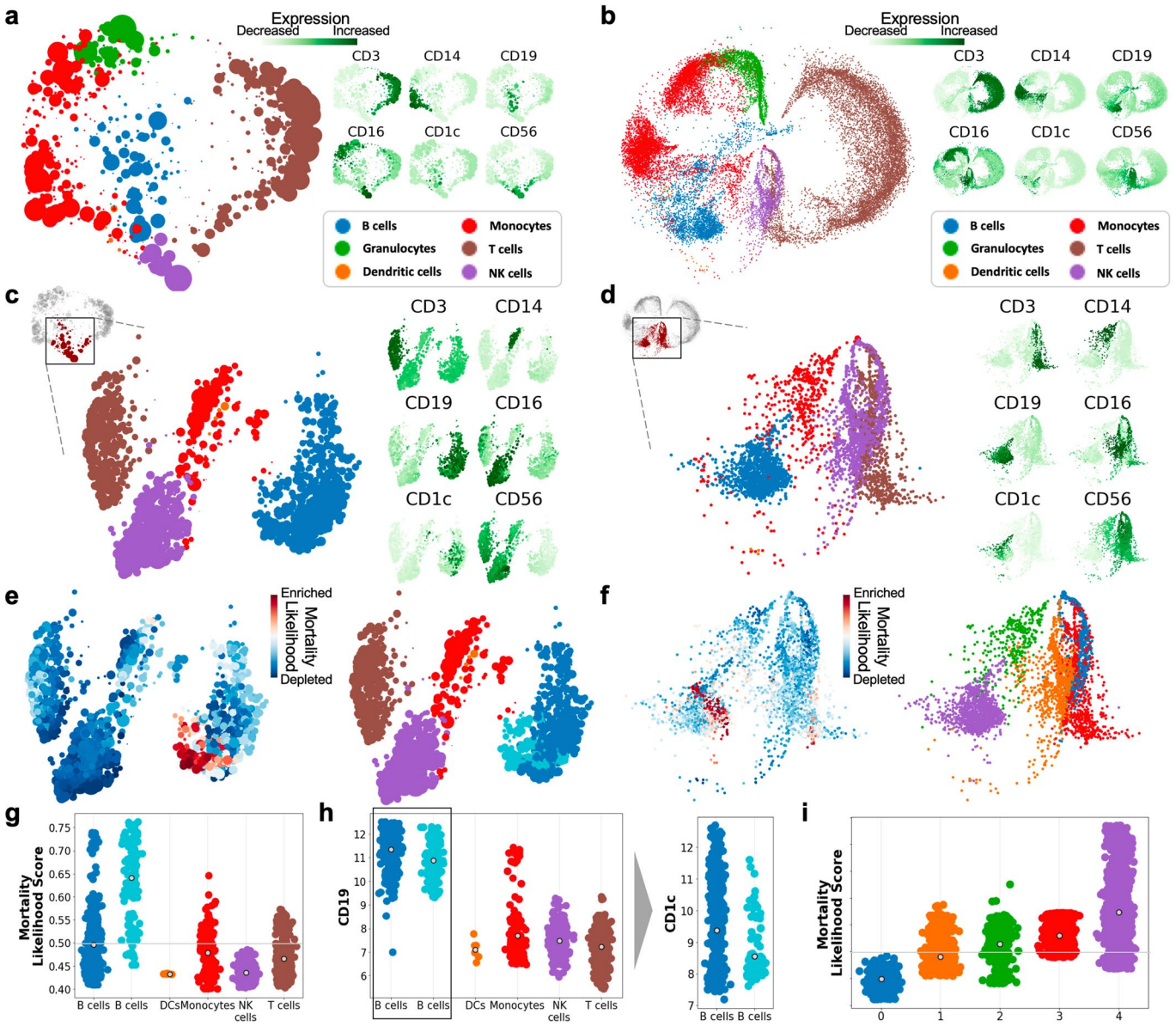
Peer review information *Nature Biotechnology* thanks Sean Bendall and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.



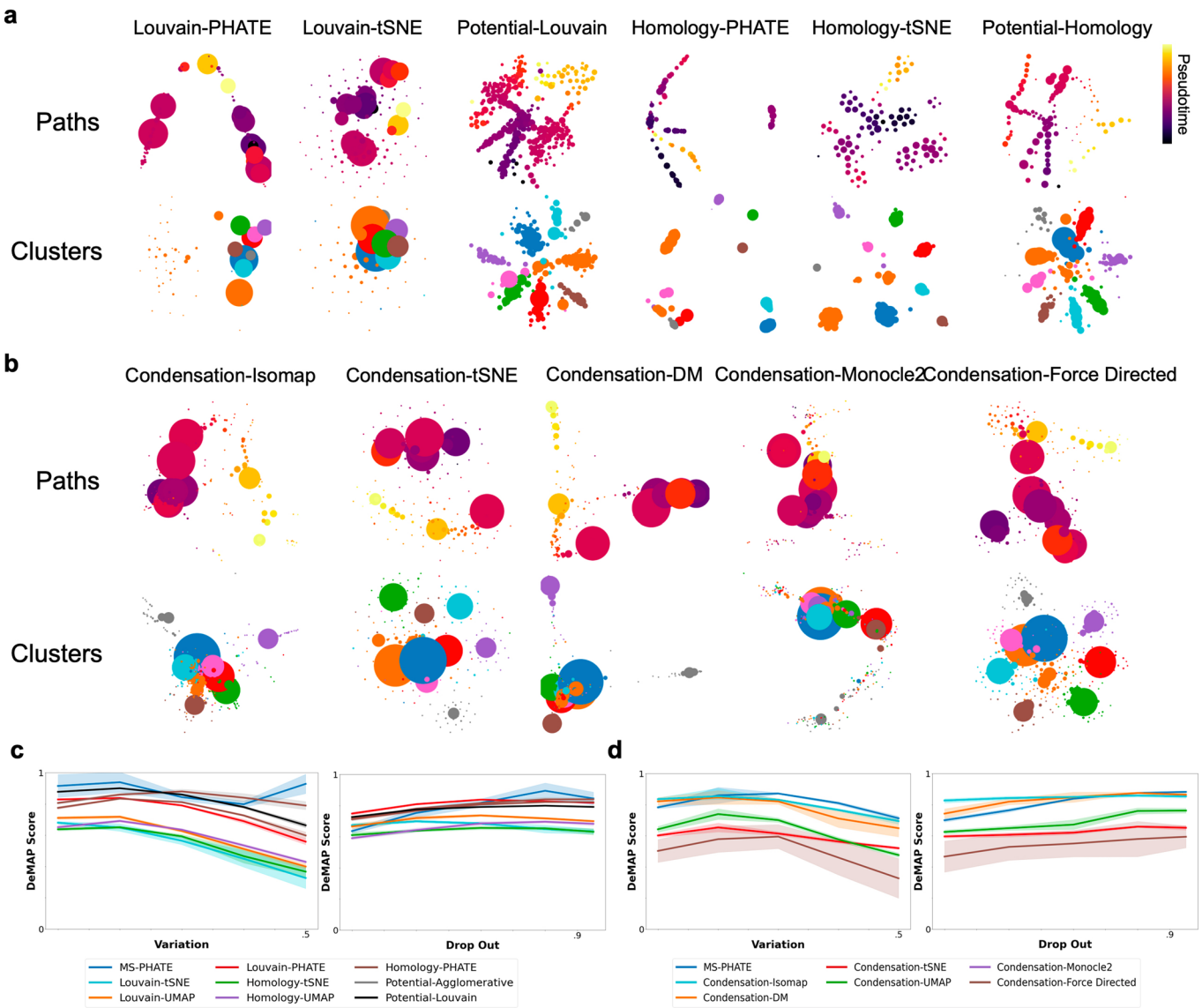
Extended Data Fig. 1 | Condensing on manifold, reproducibility and run time comparisons. a, Visualization of toy swiss roll after performing condensation in euclidean space or on diffusion potential. Top: schematic of the movement vectors of each point when run in euclidean space or on diffusion potential for one iteration. Bottom: Visualization of toy swiss roll dataset after several iterations of diffusion condensation, running in both euclidean space and diffusion potential. **b**, Comparison of diffusion condensation on diffusion potential to diffusion condensation on ambient measurement dimensions on an increasingly noisy stochastic block model to simulate nonlinear noise in a high-dimensional space. In this model, increasing amounts of Gaussian noise were added to the edge weights of the adjacency matrix. **c**, Pipeline for identifying cellular populations enriched based on clinical variables with Multiscale PHATE and MELD. **d**, Comparing run time across visualization techniques on increasingly high-dimensional flow cytometry data. **e**, Visualization of reproduciblity of Multiscale PHATE across two different runs of PBMCs measured by scRNA-seq. Each run was initialized with a different random seed.



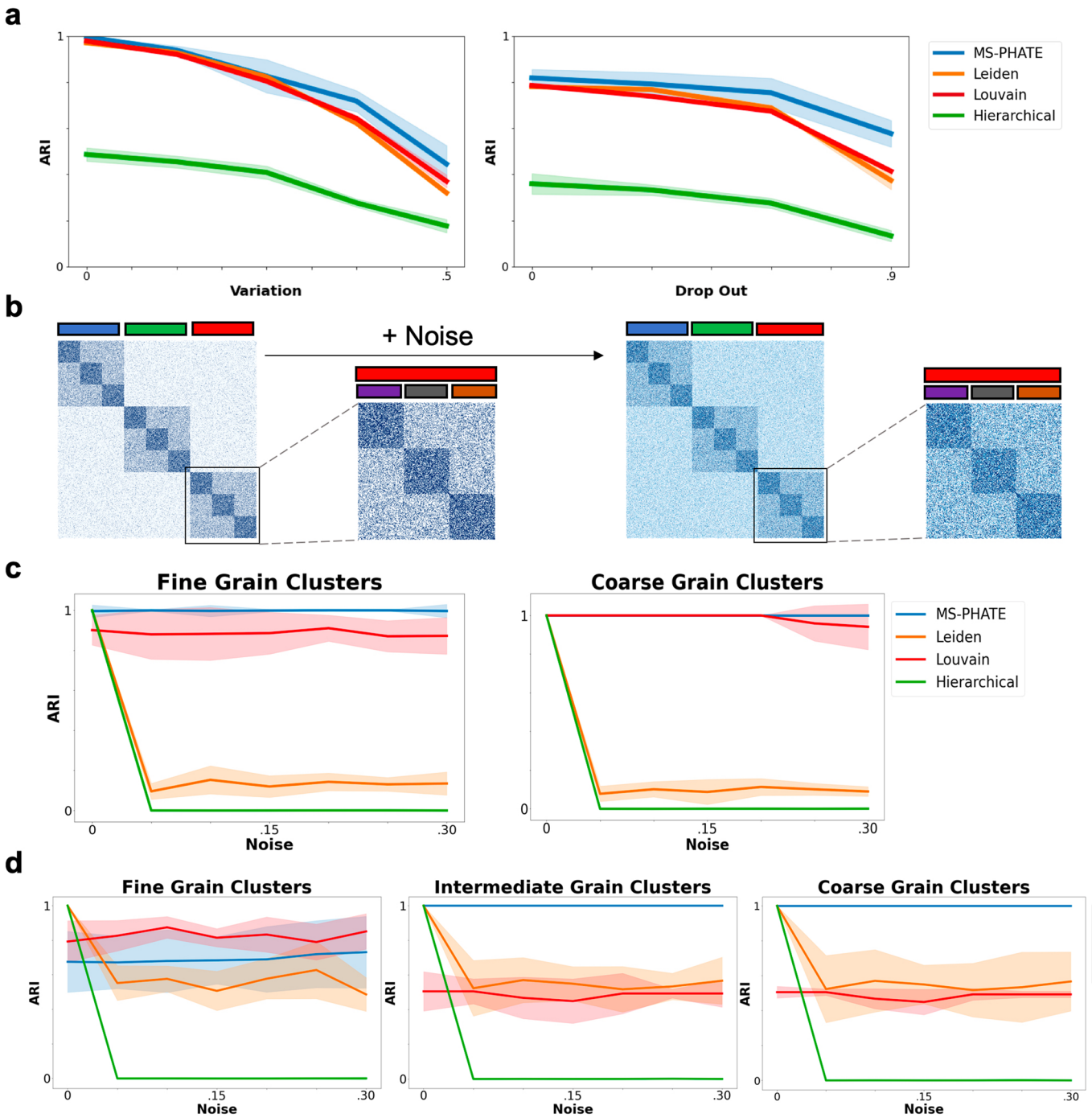
Extended Data Fig. 2 | Visualization of differing high-dimensional biological data types. Visualization comparison across a range of data types: 22 million PBMCs measured by flow cytometry (Lucas et al.), 49,942 PBMCs by scRNA-seq (Lee et al.), 2,135 patients admitted to YNHH by demographic and lab clinical variables, 25,528 cells from a diverse set of mouse tissues measured by scATAC-seq (Cusanovich et al.), 1,010,964 PBMCs measured by CyTOF (Hartmann et al.) and 50,000 TCRs from COVID-19 infected patients and healthy controls (Nolan et al., Corrie et al.).



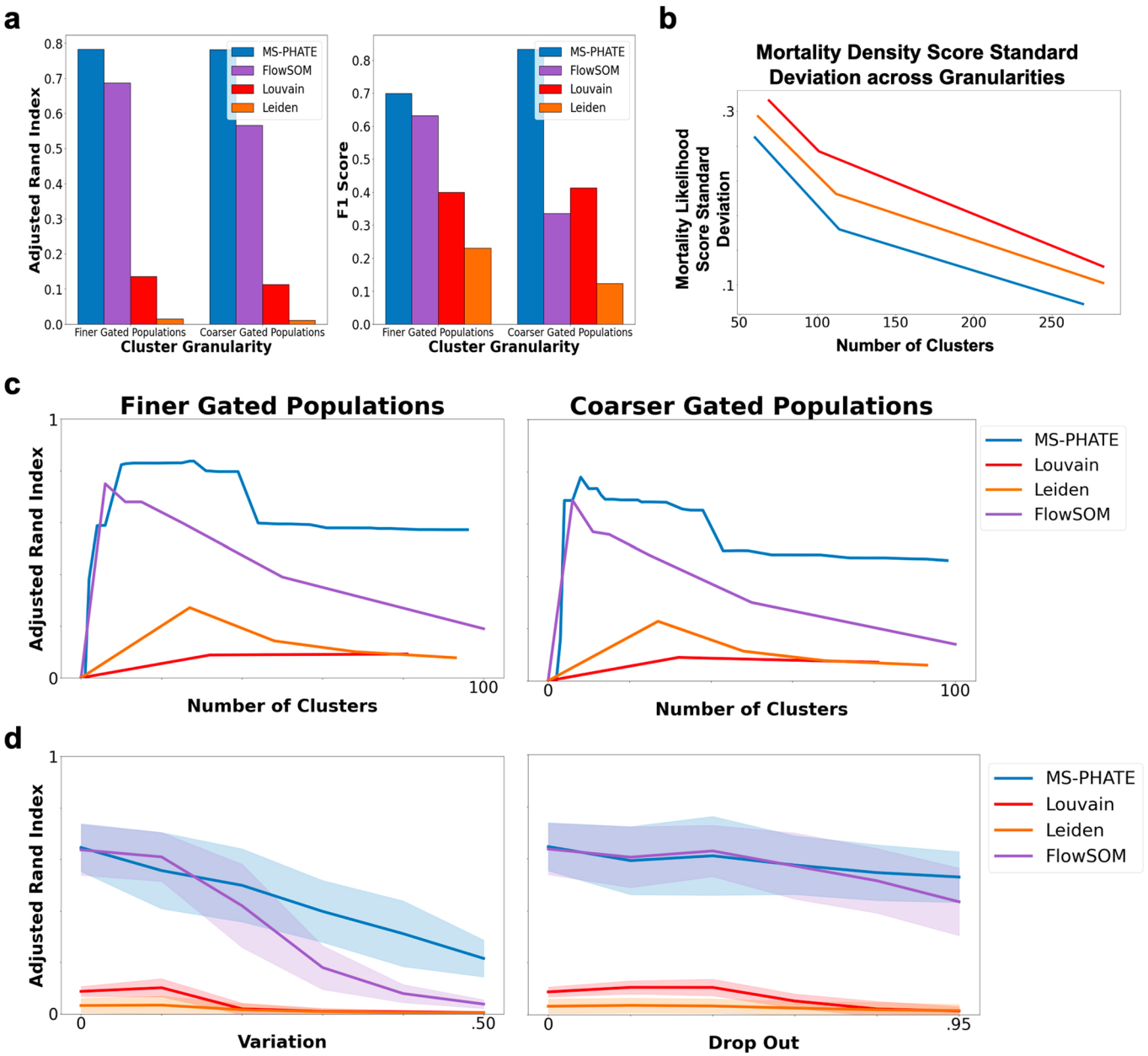
Extended Data Fig. 3 | Multiscale PHATE is capable of identify extractable cellular subsets from massive single-cell data. a, Multiscale PHATE visualization of PBMCs identifies all major cell types based on cell type-specific markers. **b**, PHATE visualization of subsample of 25,000 PBMCs helps identify all major cell types based on cell type-specific markers using Multiscale PHATE clustering. **c**, Zoom in of subsection of Multiscale PHATE manifold resolves crowding in coarse grain visualization. **d**, Zoom in of subsection of PHATE manifold does not resolve crowding. **e**, Multiscale PHATE is able to identify subpopulations enriched in patients who die from COVID. The plot on the right is colored by Multiscale PHATE-identified clusters. **f**, PHATE and vertex frequency clustering (VFC) are unable to identify subpopulations enriched in patients who die from COVID. The plot on the right is colored by VFC identified clusters. **g**, Multiscale PHATE-identified populations show differing enrichments in patients who die from COVID19. One of the B cell subsets (lighter blue color) are enriched in patients who die from COVID. **h**, Multiscale PHATE's hierarchical approach to clustering provides a gating strategy to isolate subsets of B cells enriched in patients who die from COVID19. **i**, VFC identified populations do not isolate mortality enriched cellular subsets as well as Multiscale PHATE.



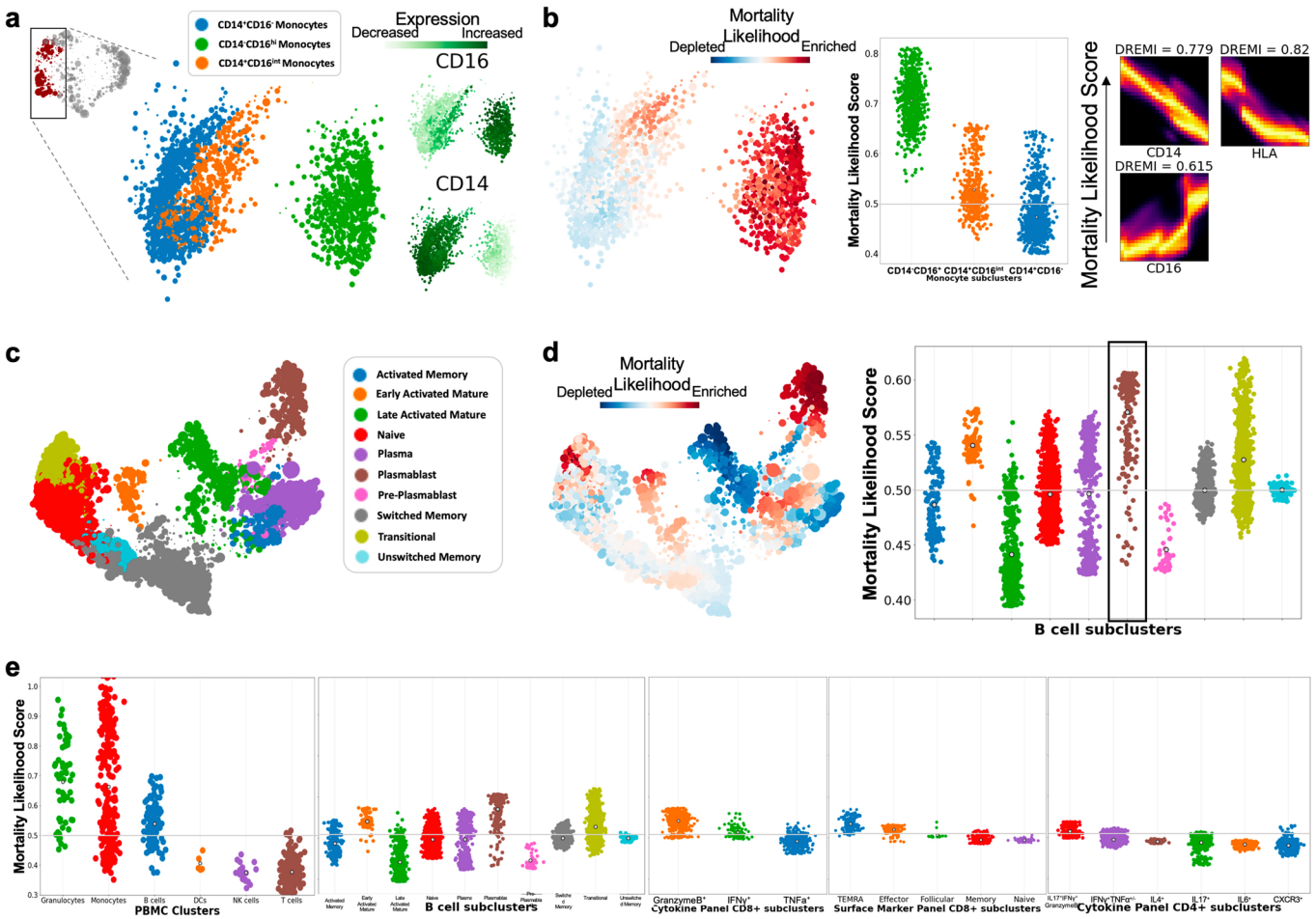
Extended Data Fig. 4 | Visualization of differing multiscale dimensionality reduction techniques. a, Visualization of noisy splatter data with either path of cluster geometry embedded with algorithms created for condensation ablation study performed in Fig. 2b. **b**, Visualization of noisy splatter data with either path of cluster geometry embedded with algorithms created for PHATE ablation study performed in Fig. 2c. **c**, Quantitative study comparing embeddings produced by Multiscale PHATE and visualization strategies which either employ community based or topologically based abstractions of data on 1.7 million cells from FlowCAP I Normal Donor (ND) dataset. Comparisons were evaluated using DeMAP with increasing levels of 2 different types of biological noise, dropout and variation. Shading represents standard deviation around mean DeMAP score for each comparison. **d**, Quantitative study comparing embeddings produced by Multiscale PHATE and visualization strategies which visualize condensation based abstractions of data. Comparisons were run and represented as described in **b**.



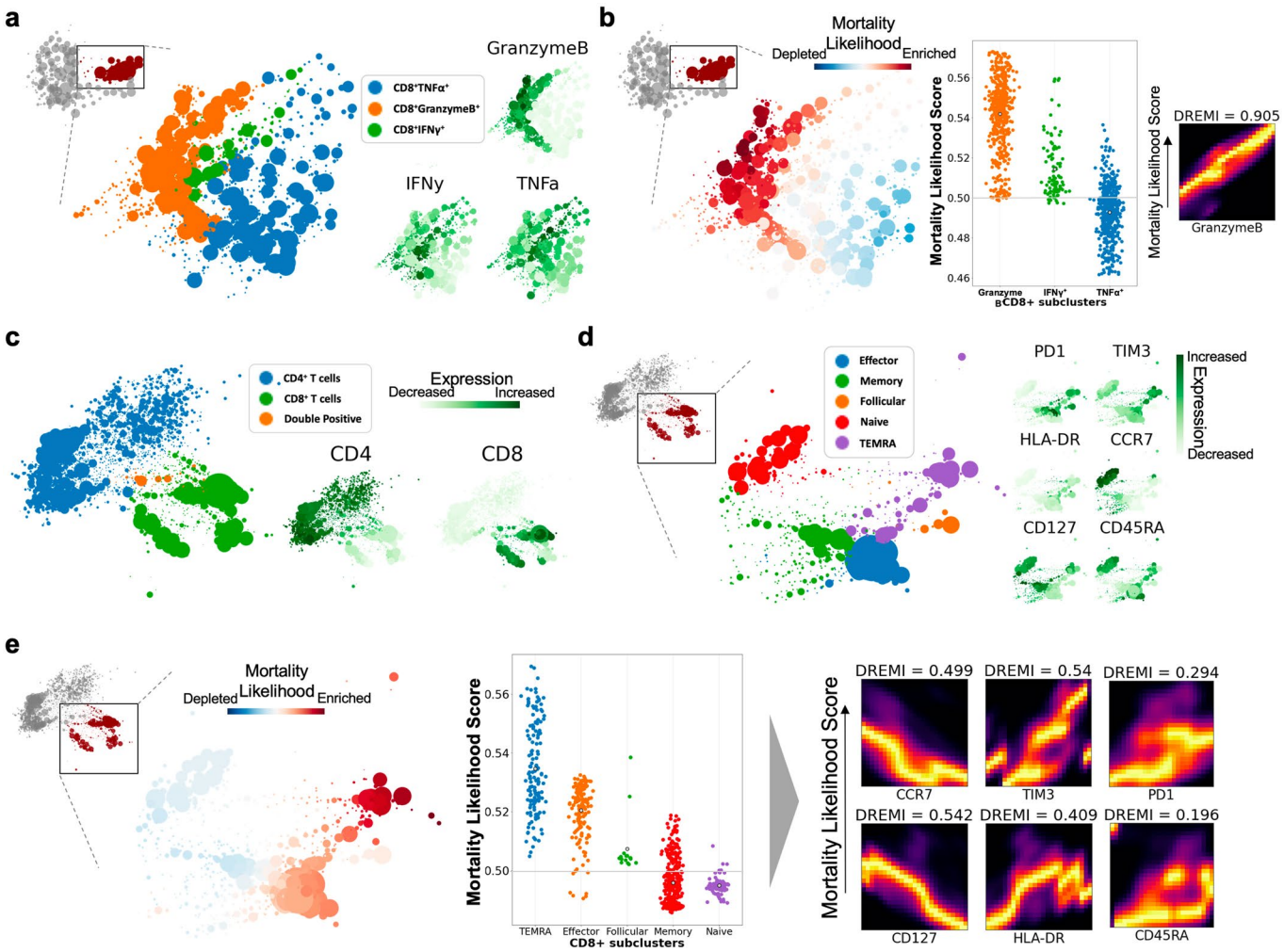
Extended Data Fig. 5 | Comparison of Multiscale PHATE with other clustering techniques on hierarchical stochastic block model. a, Computed Adjusted Rand Index (ARI) between each algorithm's predicted clusters and the known clusters on synthetic single-cell data generated by splatter (Zappia et al.) across a range of noise types, dropout and biological variation, and noise levels. Shading represents one standard deviation around mean ARI score for each comparison. **b**, Schematic of the hierarchical stochastic block model we generated for multigranular cluster comparisons. For each method, increasing amounts of random Gaussian noise were added to the adjacency matrix of stochastic block model to simulate increasing amounts of noise. While adding noise directly to data introduces simple linear noise, adding Gaussian noise to the edge weights of an adjacency matrix simulates more complex non-linear type of noise which is often present in high-dimensional biological data. **c**, 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 perturbed with increasing amounts of noise. Shading represents one standard deviation around mean ARI score for each comparison. **d**, Computed Adjusted Rand Index (ARI) between each algorithm's predicted clusters and the known clusters across coarse, intermediate and fine granularities of 3 layer stochastic block model perturbed with increasing amounts of noise. Shading represents one standard deviation around mean ARI score for each comparison.



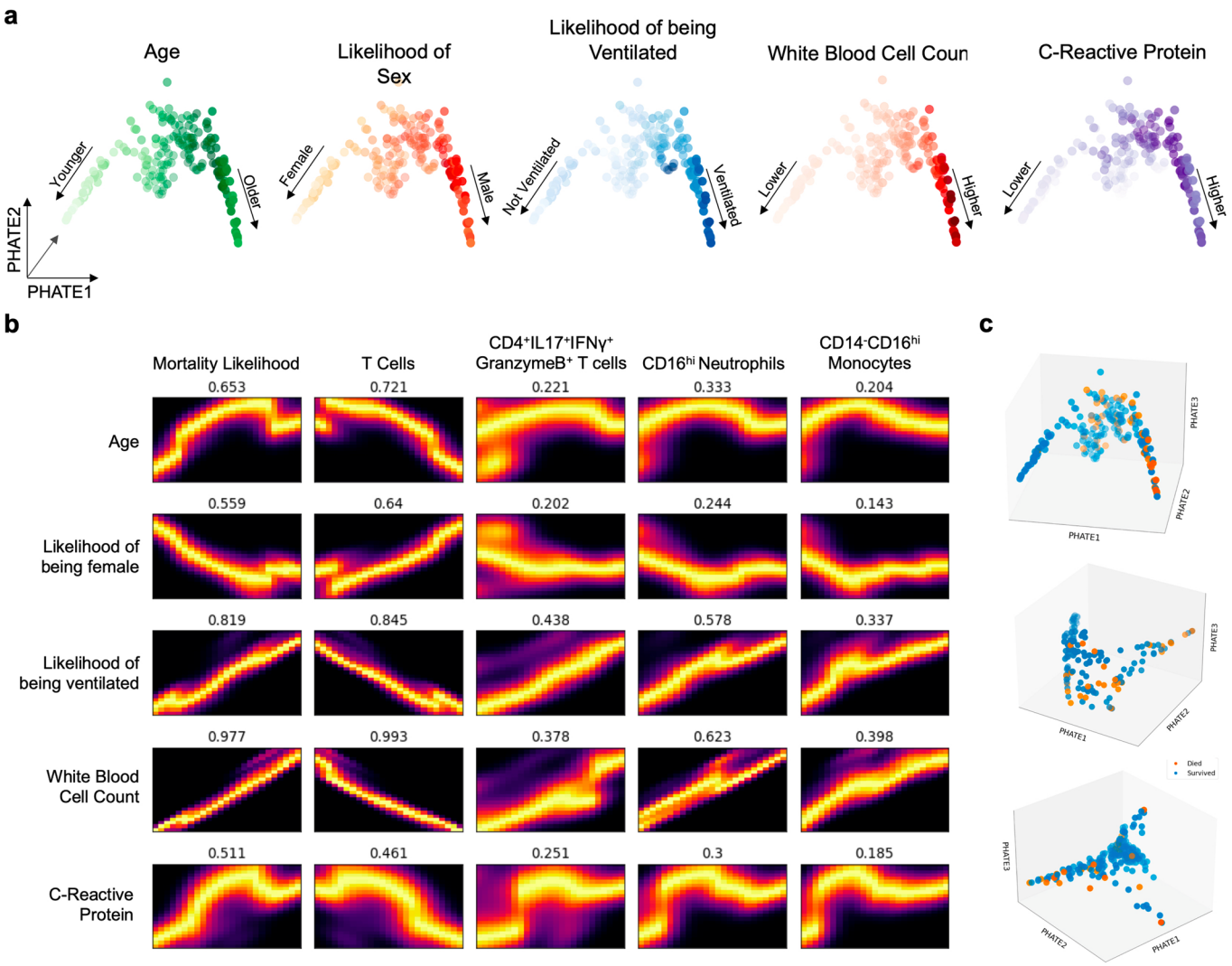
Extended Data Fig. 6 | Comparison of Multiscale PHATE with other clustering tools on real data. a, Comparison of multiple clustering approaches on flow cytometry data where cell types and subtypes have been identified through gating analysis. Clusters identified by different approaches were compared to gated populations using ARI and F1 score. **b**, Comparison of multiple clustering techniques at identifying regions with uniform MELD likelihood scores across a range of comparable granularities. **c**, Comparison of multiple clustering techniques across a range of granularities on flow cytometry data with cell types and subtypes identified as done in **a**. **d**, Comparison of multiple clustering techniques across increasing amounts of noise of different types, biological variation and dropout, as done in Extended Data Fig. 3. As done in Extended Data Fig. 3, noise was added to FlowCAP I Normal Donor (ND) dataset with known clusters. Shading represents one standard deviation around mean ARI score for each comparison.



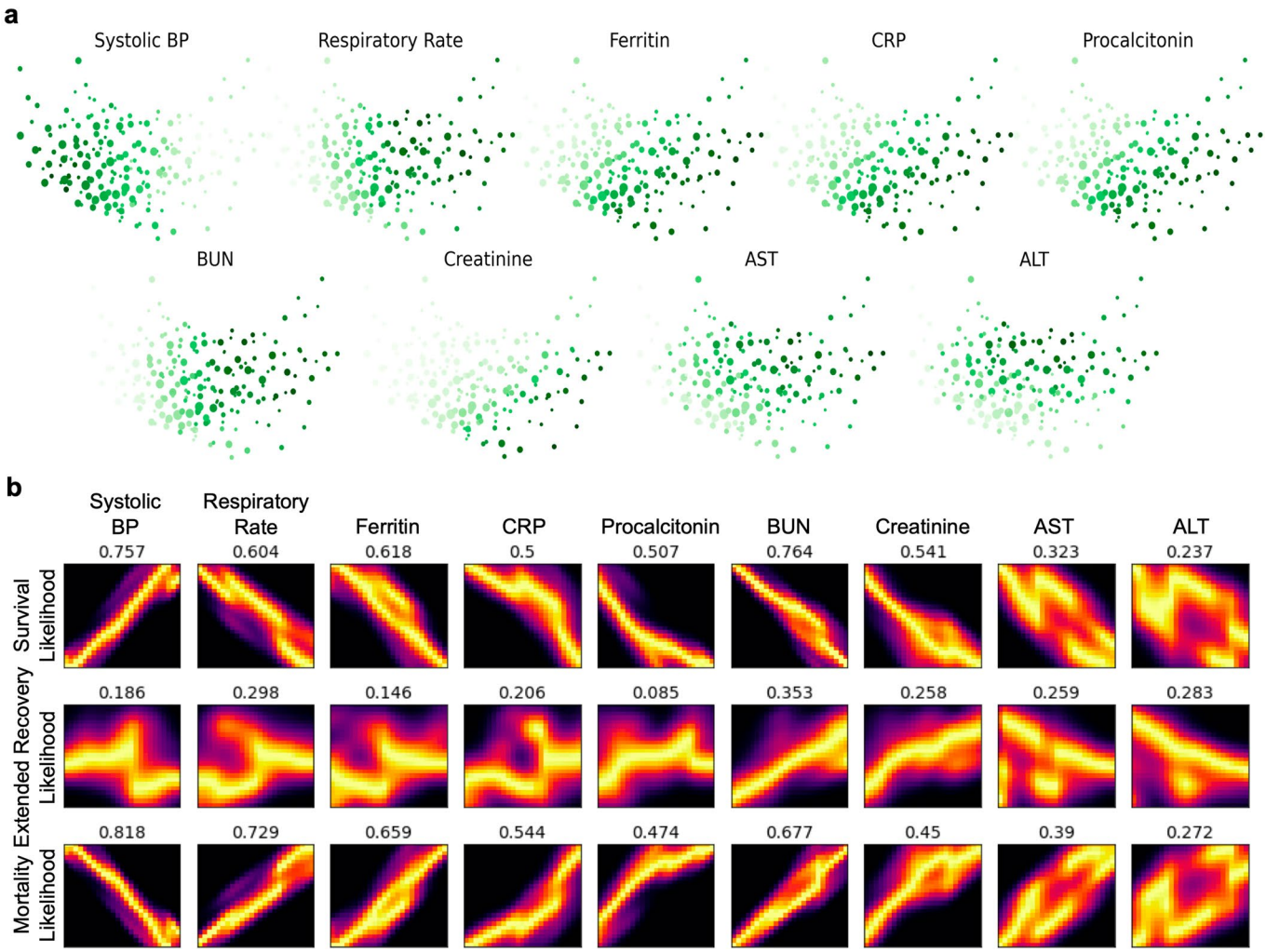
Extended Data Fig. 7 | Multiscale PHATE identifies subsets of monocytes and B cells enriched in patients who died of COVID-19. a, Zoom in of monocyte population identifies subsets based on expression of markers. Colors denote cell type and size of a dot is proportional to number of cells represented. **b**, Visualization of mortality likelihood score as computed by MELD 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. **c**, Visualization of B cells panel identifies a range of subsets based on expression of known markers. Colors denote cell type and size of a dot is proportional to number of cells represented. **d**, Visualization of mortality likelihood score identifies B cell subsets enriched in patients who die from COVID-19. **e**, 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 Fig. 8 | Multiscale PHATE analysis identifies subsets of CD8⁺ T cells enriched in patients with poor COVID-19 outcomes. a, Zoom in of CD8⁺ T cells identifies subsets based on expression of markers. Colors denote cell type and size of a dot is proportional to number of cells represented. **b**, Visualization of mortality likelihood score as computed by MELD in CD8⁺ T cells identifies subsets enriched in patients who die from COVID-19. Key associations between Granzyme B and mortality likelihood computed by DREMI and visualized with DREVI. **c**, Multiscale PHATE visualization of T cell focused surface marker panel with broad T cell subtypes identified. Colors denote cell type and size of a dot is proportional to number of cells represented. **d**, Zoom in of CD8⁺ T cells identifies subsets based on expression of known markers. **e**, Visualization of mortality likelihood score as computed by MELD in CD8⁺ 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 Fig. 9 | Visualization of patient manifold and correlation with clinical features. a, Visualizing clinical variables on patient manifold. Darker color indicates higher normalized numerical values. **b**, DREMI and DREVI association analysis between clinical variables and mortality as well as cellular populations. **c**, PHATE visualizations of patient manifolds created by Multiscale PHATE (top), conventional flow cytometry gating (middle) and single resolution of louvain clusters (bottom). Patients who died are highlighted in orange.



Extended Data Fig. 10 | Visualization of multiscale clinical manifold and correlation with patient clinical features. a, Visualizing clinical variables on clinical manifold as computed by Multiscale PHATE. Size of a dot is proportional to number of patients represented and darker color indicates higher normalized numerical values. **b**, DREMI and DREVI association analysis between clinical features and patient hospitalization outcome likelihood as computed by MELD.

nature research

Corresponding author(s): Smita Krishnaswamy

Last updated by author(s): Dec 5, 2021

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	EPIC EHR software (retrospective EMR review and clinical data aggregation) and REDCap 9.3.6 (clinical data aggregation)
Data analysis	FlowJo (version 10.6, Tree Star), GraphPad PRISM version 8.0.2 (pre-processing), Multiscale PHATE (downstream analysis)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data generated during the current study will be available in the ImmPort Platform under study number SDY1886.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was determined based on the number of patients admitted to Yale New Haven Hospital (YNHH) between March 18th and May 28th that were enrolled and consented with th current study. This study enrolled 168 patients admitted to the Yale New Haven Health care network under IRB and HIC approved protocol #2000027690. Patients were identified though screening of EMR records for potential enrollment. Informed consent was obtained by trained staff and sample collection commenced immediately upon study enrollment.
Data exclusions	168 COVID-19 patients were enrolled on this study however 37 were excluded. Those included: Pregnant women and patients on active chemotherapy.
Replication	The findings were not replicated.
Randomization	Patients were stratified by disease severity (moderate and severe) based on disease outcome (death or discharge respectively).
Blinding	At the time of sample acquisition and processing, scientists were unaware of the patients' condition and severity. Blood acquisition is performed and recorded by a separate team. Information of patients' conditions are not available until after processing and preliminary analysis with Multiscale PHATE. At this time, outcome was unblinded to allow for clinical correlation.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☐ Antibodies
☐	☐ Eukaryotic cell lines
☐	☐ Palaeontology and archaeology
☐	☐ Animals and other organisms
☐	☒ Human research participants
☐	☐ Clinical data
☐	☐ Dual use research of concern

Methods

n/a	Involved in the study
☐	☐ ChIP-seq
☐	☒ Flow cytometry
☐	☐ MRI-based neuroimaging

Antibodies

Antibodies used

All antibodies used in this study are against human proteins. BB515 anti-hHLA-DR (G46-6) (1:400) (BD Biosciences), BV785 antihCD16 (3G8) (1:100) (BioLegend), PE-Cy7 anti-hCD14 (HCD14) (1:300) (BioLegend), BV605 anti-hCD3 (UCHT1) (1:300) (BioLegend), BV711 anti-hCD19 (SJ25C1) (1:300) (BD Biosciences), AlexaFluor647 anti-hCD1c (L161) (1:150) (BioLegend), Biotin anti-hCD141 (M80) (1:150) (BioLegend), PE-Dazzle594 anti-hCD56 (HCD56) (1:300) (BioLegend), PE anti-hCD304 (12C2) (1:300) (BioLegend), APCFire750 anti-hCD11b (ICRF44) (1:100) (BioLegend), PerCP/Cy5.5 anti-hCD66b (G10F5) (1:200) (BD Biosciences), BV785 anti-hCD4 (SK3) (1:200) (BioLegend), APCFire750 or PE-Cy7 or BV711 anti-hCD8 (SK1) (1:200) (BioLegend), BV421 anti-hCCR7 (G043H7) (1:50) (BioLegend), AlexaFluor 700 anti-hCD45RA (HI100) (1:200) (BD Biosciences), PE anti-hPD1 (EH12.2H7) (1:200) (BioLegend), APC anti-hTIM3 (F38-2E2) (1:50) (BioLegend), BV711 anti-hCD38 (HIT2) (1:200) (BioLegend), BB700 anti-hCXCR5 (RF8B2) (1:50) (BD Biosciences), PECy7 anti-hCD127 (HIL-7R-M21) (1:50) (BioLegend), PE-CF594 anti-hCD25 (BC96) (1:200) (BD Biosciences), BV711 anti-hCD127 (HIL-7R-M21) (1:50) (BD Biosciences), BV421 anti-hIL17a (N49-653) (1:100) (BD Biosciences), AlexaFluor 700 anti-hTNFa (MAb11) (1:100) (BioLegend), PE or APC/Fire750 anti-hIFNy (4S.B3) (1:60) (BioLegend), FITC anti-hGranzymeB (GB11) (1:200) (BioLegend), AlexaFluor 647 anti-hIL-4 (8D4-8) (1:100) (BioLegend), BB700 anti-hCD183/CXCR3 (1C6/CXCR3) (1:100) (BD Biosciences), PE-Cy7 antihIL-6 (MQ2-13A5) (1:50) (BioLegend), PE anti-hIL-2 (5344.111) (1:50) (BD Biosciences), BV785 anti-hCD19 (SJ25C1) (1:300) (BioLegend), BV421 anti-hCD138 (MI15) (1:300) (BioLegend), AlexaFluor700 anti-hCD20 (2H7) (1:200) (BioLegend), AlexaFluor 647 anti-hCD27 (M-T271) (1:350) (BioLegend), PE/Dazzle594 anti-hIgD (IA6-2) (1:400) (BioLegend), PE-Cy7 anti-hCD86 (IT2.2) (1:100) (BioLegend), APC/Fire750 anti-hIgM (MHM-88) (1:250) (BioLegend), BV605 anti-hCD24 (ML5) (1:200) (BioLegend), BV421 anti-hCD10 (HI10a) (1:200) (BioLegend), BV421 anti-CDh15 (SSEA-1) (1:200) (BioLegend), AlexaFluor 700 Streptavidin (1:300) (ThermoFisher), BV605 Streptavidin (1:300) (BioLegend).

Validation

All antibodies used in this study are commercially available, and all have been validated by the manufacturers and used by other publications. Likewise, we titrated these antibodies according to our own our staining conditions. The following were validated in the following species: BB515 anti-hHLA-DR (G46-6) (BD Biosciences) (Human, Rhesus, Cynomolgus, Baboon), BV785 anti-hCD16 (3G8) (BioLegend) (Human, African Green, Baboon, Capuchin Monkey, Chimpanzee, Cynomolgus, Marmoset, Pigtailed Macaque, Rhesus, Sooty Mangabey, Squirrel Monkey), PE-Cy7 anti-hCD14 (HCD14) (BioLegend) (Human), BV605 anti-hCD3 (UCHT1) (BioLegend) (Human, Chimpanzee), BV711 anti-hCD19 (SJ25C1) (BD Biosciences) (Human), AlexaFluor647 anti-hCD1c (L161) (BioLegend) (Human, African Green, Baboon, Cynomolgus, Rhesus), Biotin anti-hCD141 (M80) (BioLegend) (Human, African Green, Baboon), PE-Dazzle594 anti-hCD56 (HCD56) (BioLegend) (Human, African Green, Baboon, Cynomolgus, Rhesus), PE anti-hCD304 (12C2) (BioLegend) (Human), APCFire750 anti-hCD11b (ICRF44) (BioLegend) (Human, African Green, Baboon, Chimpanzee, Common Marmoset, Cynomolgus, Rhesus, Swine), PerCP/Cy5.5 anti-hCD66b (G10F5) (BD Biosciences) (Human), BV785 anti-hCD4 (SK3) (BioLegend) (Human), APCFire750 or PE-Cy7 or BV711 anti-hCD8 (SK1) (BioLegend) (Human, Cross-Reactivity: African Green, Chimpanzee, Cynomolgus, Pigtailed Macaque, Rhesus, Sooty Mangabey), BV421 anti-hCCR7 (G043H7) (BioLegend) (Human, African Green, Baboon, Cynomolgus, Rhesus), AlexaFluor 700 anti-hCD45RA (HI100) (BD Biosciences) (Human), PE anti-hPD1 (EH12.2H7) (BioLegend) (Human, African Green, Baboon, Chimpanzee, Common Marmoset, Cynomolgus, Rhesus, Squirrel Monkey), APC antihTIM3 (F38-2E2) (BioLegend) (Human), BV711 anti-hCD38 (HIT2) (BioLegend) (Human, Chimpanzee, Horse), BB700 anti-hCXCR5 (RF8B2) (BD Biosciences) (Human), PE-Cy7 anti-hCD127 (HIL-7R-M21) (BioLegend) (Human), PE-CF594 anti-hCD25 (BC96) (BD Biosciences) (Human, Rhesus, Cynomolgus, Baboon), BV711 anti-hCD127 (HIL-7R-M21) (BD Biosciences) (Human), BV421 anti-hIL-17a (N49-653) (BD Biosciences) (Human), AlexaFluor 700 anti-hTNFa (MAb11) (BioLegend) (Human, Cat, Cross-Reactivity: Chimpanzee, Baboon, Cynomolgus, Rhesus, Pigtailed Macaque, Sooty Mangabey, Swine), PE or APC/Fire750 anti-hIFNy (4S.B3) (BioLegend) (Human, Cross-Reactivity: Chimpanzee, Baboon, Cynomolgus, Rhesus), FITC anti-hGranzymeB (GB11) (BioLegend) (Human, Mouse, Cross-Reactivity: Rat), AlexaFluor 647 anti-hIL-4 (8D4-8) (BioLegend) (Human, Cross-Reactivity: Chimpanzee, Baboon, Cynomolgus, Rhesus), BB700 anti-hCD183/CXCR3 (1C6/CXCR3) (BD Biosciences) (Human, Rhesus, Cynomolgus, Baboon), PE-Cy7 anti-IL-6 (MQ2-13A5) (BioLegend) (Human), PE anti-hIL-2 (5344.111) (BD Biosciences) (Human), BV785 anti-hCD19 (SJ25C1) (BioLegend) (Human), BV421 anti-hCD138 (MI15) (BioLegend) (Human), AlexaFluor700 anti-hCD20 (2H7) (BioLegend) (Human, Baboon, Capuchin Monkey, Chimpanzee, Cynomolgus, Pigtailed Macaque, Rhesus, Squirrel Monkey), AlexaFluor 647 anti-hCD27 (M-T271) (BioLegend) (Human, Cross-Reacitivity: Baboon, Cynomolgus, Rhesus), PE/Dazzle594 anti-hIgD (IA6-2) (BioLegend) (Human), PE-Cy7 anti-hCD86 (IT2.2) (BioLegend) (Human, African Green, Baboon, Capuchin Monkey, Common Marmoset, Cotton-topped Tamarin, Chimpanzee, Cynomolgus, Rhesus), APC/Fire750 anti-hIgM (MHM-88) (BioLegend) (Human, African Green, Baboon, Cynomolgus, Rhesus), BV605 anti-hCD24 (ML5) (BioLegend) (Human, Cross-Reactivity: Chimpanzee), BV421 anti-hCD10 (HI10a) (BioLegend) (Human, African Green, Baboon, Capuchin monkey, Chimpanzee, Cynomolgus, Rhesus), BV421 anti-hCD15 (SSEA-1) (BioLegend) (Human), AlexaFluor 700 Streptavidin (1:300) (ThermoFisher), BV605 Streptavidin (1:300) (BioLegend).

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

State the source of each cell line used.

Authentication

Describe the authentication procedures for each cell line used OR declare that none of the cell lines used were authenticated.

Mycoplasma contamination

Confirm that all cell lines tested negative for mycoplasma contamination OR describe the results of the testing for mycoplasma contamination OR declare that the cell lines were not tested for mycoplasma contamination.

Commonly misidentified lines
(See ICLAC register)

Name any commonly misidentified cell lines used in the study and provide a rationale for their use.

Palaeontology and Archaeology

Specimen provenance

Provide provenance information for specimens and describe permits that were obtained for the work (including the name of the issuing authority, the date of issue, and any identifying information).

Specimen deposition

Indicate where the specimens have been deposited to permit free access by other researchers.

Dating methods

If new dates are provided, describe how they were obtained (e.g. collection, storage, sample pretreatment and measurement), where they were obtained (i.e. lab name), the calibration program and the protocol for quality assurance OR state that no new dates are provided.

☐ Tick this box to confirm that the raw and calibrated dates are available in the paper or in Supplementary Information.

Ethics oversight

Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

For laboratory animals, report species, strain, sex and age OR state that the study did not involve laboratory animals.

Wild animals

Provide details on animals observed in or captured in the field; report species, sex and age where possible. Describe how animals were caught and transported and what happened to captive animals after the study (if killed, explain why and describe method; if released, say where and when) OR state that the study did not involve wild animals.

Field-collected samples	*For laboratory work with field-collected samples, describe all relevant parameters such as housing, maintenance, temperature, photoperiod and end-of-experiment protocol OR state that the study did not involve samples collected from the field.*
Ethics oversight	*Identify the organization(s) that approved or provided guidance on the study protocol, OR state that no ethical approval or guidance was required and explain why not.*

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Cohort characteristics: 63.52 mean age; 31.24 mean BMI; 48% female; 32% Black-African American, 52% white, 16% hispanic, 2% unknown.
Recruitment	Patients admitted to the Yale New Haven Hospital (YNHH) between the 18th of March through the 28th of 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. (serology was further confirmed for all patients enrolled). Patients were identified though screening of EMR records for potential enrollment with no self selection. Informed consent was obtained by trained staff and sample collection commenced immediately upon study enrollment.
Ethics oversight	Yale Human Research Protection Program Institutional Review Boards. Informed consents were obtained from all enrolled patients. Our research protocol was reviewed and approved by the Yale School of Medicine IRB and HIC (#2000027690). Informed consent was obtained by trained staff and records maintained in our research database for the duration of our study. There were no minors included on this study.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	*Provide the trial registration number from ClinicalTrials.gov or an equivalent agency.*
Study protocol	*Note where the full trial protocol can be accessed OR if not available, explain why.*
Data collection	*Describe the settings and locales of data collection, noting the time periods of recruitment and data collection.*
Outcomes	*Describe how you pre-defined primary and secondary outcome measures and how you assessed these measures.*

Dual use research of concern

Policy information about dual use research of concern

Hazards

Could the accidental, deliberate or reckless misuse of agents or technologies generated in the work, or the application of information presented in the manuscript, pose a threat to:

No	Yes	
☒	☐	Public health
☒	☐	National security
☒	☐	Crops and/or livestock
☒	☐	Ecosystems
☒	☐	Any other significant area

Experiments of concern

Does the work involve any of these experiments of concern:

No	Yes	
☐	☐	Demonstrate how to render a vaccine ineffective
☐	☐	Confer resistance to therapeutically useful antibiotics or antiviral agents
☐	☐	Enhance the virulence of a pathogen or render a nonpathogen virulent
☐	☐	Increase transmissibility of a pathogen
☐	☐	Alter the host range of a pathogen
☐	☐	Enable evasion of diagnostic/detection modalities
☐	☐	Enable the weaponization of a biological agent or toxin
☐	☐	Any other potentially harmful combination of experiments and agents

ChIP-seq

Data deposition

☐ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☐ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*
Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☒ All plots are contour plots with outliers or pseudocolor plots.

☒ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Freshly isolated PBMCs were stained for live and dead markers, blocked with Human TruStain FcX , stained for surface markers and then fixed with PFA 4%. For intracellular cytokine staining following stimulation, cells were surface stained, washed and fixed in 4% PFA. After permeabilization with 1X Permeabilization Buffer cells were stained for intracellular cytokines analysis.

Instrument	Cells were acquired on an Attune NXT (ThermoFisher).
Software	Data were analysed using FlowJo software version 10.6 software (Tree Star).
Cell population abundance	Cell population abundance: Cells populations were reported in various formats including as a number or concentration of the patient's blood sample (x106cells/mL), as a proportion of live, single PBMC (% of Live), or as a proportion of a parent gate (% of CD4 T cells, % of Monocytes, etc.). The full gating path for clarification is included in the extended figures.
Gating strategy	SSC-A and FSC-A parameters were used to select leukocytes from isolated PBMCs. Live and dead cells were defined based on aqua staining. Singlets were separated based on SSC/ FSC parameters.

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Magnetic resonance imaging

Experimental design

Design type	*Indicate task or resting state; event-related or block design.*
Design specifications	*Specify the number of blocks, trials or experimental units per session and/or subject, and specify the length of each trial or block (if trials are blocked) and interval between trials.*
Behavioral performance measures	*State number and/or type of variables recorded (e.g. correct button press, response time) and what statistics were used to establish that the subjects were performing the task as expected (e.g. mean, range, and/or standard deviation across subjects).*

Acquisition

Imaging type(s)	*Specify: functional, structural, diffusion, perfusion.*
Field strength	*Specify in Tesla*
Sequence & imaging parameters	*Specify the pulse sequence type (gradient echo, spin echo, etc.), imaging type (EPI, spiral, etc.), field of view, matrix size, slice thickness, orientation and TE/TR/flip angle.*
Area of acquisition	*State whether a whole brain scan was used OR define the area of acquisition, describing how the region was determined.*

Diffusion MRI ☐ Used ☐ Not used

Preprocessing

Preprocessing software	*Provide detail on software version and revision number and on specific parameters (model/functions, brain extraction, segmentation, smoothing kernel size, etc.).*
Normalization	*If data were normalized/standardized, describe the approach(es): specify linear or non-linear and define image types used for transformation OR indicate that data were not normalized and explain rationale for lack of normalization.*
Normalization template	*Describe the template used for normalization/transformation, specifying subject space or group standardized space (e.g. original Talairach, MNI305, ICBM152) OR indicate that the data were not normalized.*
Noise and artifact removal	*Describe your procedure(s) for artifact and structured noise removal, specifying motion parameters, tissue signals and physiological signals (heart rate, respiration).*
Volume censoring	*Define your software and/or method and criteria for volume censoring, and state the extent of such censoring.*

Statistical modeling & inference

Model type and settings	*Specify type (mass univariate, multivariate, RSA, predictive, etc.) and describe essential details of the model at the first and second levels (e.g. fixed, random or mixed effects; drift or auto-correlation).*
Effect(s) tested	*Define precise effect in terms of the task or stimulus conditions instead of psychological concepts and indicate whether ANOVA or factorial designs were used.*

Specify type of analysis: ☐ Whole brain ☐ ROI-based ☐ Both

Statistic type for inference (See Eklund et al. 2016)	*Specify voxel-wise or cluster-wise and report all relevant parameters for cluster-wise methods.*
Correction	*Describe the type of correction and how it is obtained for multiple comparisons (e.g. FWE, FDR, permutation or Monte Carlo).*

Models & analysis

n/a	Involved in the study
☐	☐ Functional and/or effective connectivity
☐	☐ Graph analysis
☐	☐ Multivariate modeling or predictive analysis

Functional and/or effective connectivity

Report the measures of dependence used and the model details (e.g. Pearson correlation, partial correlation, mutual information).

Graph analysis

Report the dependent variable and connectivity measure, specifying weighted graph or binarized graph, subject- or group-level, and the global and/or node summaries used (e.g. clustering coefficient, efficiency, etc.).

Multivariate modeling and predictive analysis

Specify independent variables, features extraction and dimension reduction, model, training and evaluation metrics.