Learning multi-cellular representations of single-cell transcriptomics data enables characterization of patient-level disease states

Tianyu Liu[1,2,*], **Edward De Brouwer**[1,*], **Tony Kuo**[1,3], **Nathaniel Diamant**[1], **Alsu Missarova**[1], **Hanchen Wang**[1,4], **Minsheng Hao**[1], **Hector Corrada Bravo**[1], **Gabriele Scalia**[1,✉], **Aviv Regev**[1,✉], **Graham Heimberg**[1,✉]

[1]Research & Early Development, Genentech, South San Francisco, 94080, CA, USA
[2]Interdepartmental Program in Computational Biology & Bioinformatics, Yale University, New Haven, 06511, CT, USA
[3]Roche Informatics, F. Hoffmann-La Roche Ltd., Mississauga, Canada
[4]Department of Computer Science, Stanford University, Palo Alto, 94035, CA, USA

Abstract

Single-cell RNA-seq (scRNA-seq) has become a prominent tool for studying human biology and disease. The availability of massive scRNA-seq datasets and advanced machine learning techniques has recently driven the development of single-cell foundation models that provide informative and versatile cell representations based on expression profiles. However, to understand disease states, we need to consider entire tissue ecosystems, simultaneously considering many different interacting cells. Here, we tackle this challenge by generating *patient-level* representations derived from multi-cellular expression context measured with scRNA-seq of tissues. We develop PaSCient, a novel model that employs a multi-level representation learning paradigm and provides importance scores at the individual cell and gene levels for fine-grained analysis across multiple cell types and gene programs characteristic of a given disease. We apply PaSCient to learn a disease model across a large-scale scRNA-seq atlas of 24.3 million cells from over 5,000 patients. Comprehensive and rigorous benchmarking demonstrates the superiority of PaSCient in disease classification and its multiple downstream applications, including dimensionality reduction, gene/cell type prioritization, and patient subgroup discovery.

1 Introduction

Technological innovations in the past decade have led to the collection of vast and exponentially growing amounts of data for biological research, which can help revolutionize our understanding of human disease biology Dash et al. (2019); Arowoogun et al. (2024); Obermeyer & Emanuel (2016); Marx (2013). In particular, the advent of single-cell RNA-seq (scRNA-seq) has enabled the charting of the heterogeneity of cell states and functions, by profiling the expression of hundreds of millions of cells Regev et al. (2017). The large number of cell profiles within and across experiments has opened the way to discoveries from new cell types Jindal et al. (2018), distinct genes programs associated with response to therapy or drug resistance, specific marker genes Pullin & McCarthy (2024);

* Both authors contributed equally to this work.
✉ Corresponding authors. Emails: heimberg@gene.com, regev.aviv@gene.com, scaliag@gene.com.

AI for New Drug Modalities at NeurIPS 2024.

Liu et al. (2024), and unique patient subsets Rood et al. (2022); Suvà et al. (2014). Nevertheless, most scRNA-seq studies were analyzed in isolation and only from a limited number of patients, hindering our ability to understand biological processes at a patient level Kau & Korenblat (2014); Schaid et al. (2018); Hong et al. (2012); Ioannidis (2007); Dattani et al. (2022); pol (2021). Moreover, studies have typically focused on partitioning cells into categories (types, subtypes, states, etc) and then studying each of them separately, with only limited efforts focused on the overall ecosystem of cells assembled together. Yet, diseases typically involve breakdown of homeostasis in tissue, impacting multiple cells.

Fortunately, the growing number of scRNA-seq studies has now reached a total number of patients that can realistically support machine learning approaches capable of modeling disease biology at a patient level Barrett et al. (2012); Biology et al. (2023). Reasoning about the disease process at the patient level with the granularity of single-cell expression could potentially help uncover subgroups within patient populations (endotypes), understand or predict patient responses to therapies, and advance toward more precise and personalized medicine.

These considerations have motivated the development of machine learning models to aggregate cells to identify disease states. However, existing models only focus on binary disease classification, and were trained with only few samples and studies He et al. (2021); Mao et al. (2024); Xiong et al. (2023); Mitchel et al. (2024), failing to leverage the large repositories of single-cell expression data available. A more recent work incorporates a larger patient corpus but focuses on multi-modal biomedical data integration, and limits its disease prediction to COVID-19 only Litinetskaya et al. (2024). By contrast, we aspire to a method that can leverage the full scope of available data and jointly model all diseases in a single model. However, this vision comes with significant challenges, such as the inherent confounding and batch effects of pooling together data from different studies Leek et al. (2010), the imbalanced composition of different tissues, cell types, and diseases Ferretti et al. (2018), and the noise of scRNA-seq data Janssen et al. (2023); Chu et al. (2022).

Here, we propose PaSCient, a foundation model that produces a patient representation based on the gene expression of all cells in a patient's sample, by leveraging large scale single-cell expression studies across different tissues and disease. Intuitively, each patient is represented as a set (or bag) of cells, which our model processes to provide a biologically informed vector representation of the patient. To achieve patient-level representations, we rely on a dedicated attention-based aggregation mechanism and data resampling strategy, which addresses the data integration challenges Wang et al. (2024); Boyeau et al. (2022) posed by the dataset heterogeneity. Our versatile representation can then be used to compare, cluster, or classify patients. To elucidate disease mechanisms at the patient level, we propose an interpretable mechanism based on integrated gradients Sundararajan et al. (2017) to score individual genes and/or cell types in a given patient prediction. This enables a remarkably fine-grained gene or cell-type prioritization, supporting biological discovery at the patient level in terms of individual genes, specific cell types, multiple cell types (simultaneously) and their interconnections. Our comprehensive and rigorous benchmarking further demonstrates the superiority of PaSCient in disease classification compared to single-cell foundation models and underscores its multiple downstream applications, including dimensionality reduction, biological prioritization, and patient subgroup discovery.

To summarize, our contributions are:

1. We propose a machine-learning model that creates patient-level representations based on their single-cell expression profiles. This representation can be used to compare, cluster, or classify patients. Our model leverages single-cell expression studies from over 5,000 patients.

2. The predictions of PaSCient can be interpreted to enable fine-grained prioritization of genes, cell-types, and sets of cell types (and their genes), thereby holistically interrogating disease mechanisms at the patient level.

3. We demonstrate the capabilities of PaSCient on a COVID-19 case study, showing that the model can be used to infer disease severity subgroups and prioritize cell-type specific genes associated with the disease.

Our code is available at `https://github.com/genentech/pascient`.

2 Results

2.1 Overview of PaSCient

PaSCient takes the expression profiles of individual cells present within a patient's sample as input and produces a summarized vector representation of the patient. This representation can then be used for downstream tasks such as dimensionality reduction and visualization, biological feature prioritization, treatment response prediction, and disease severity prediction, among others (Figure 1(a)).

Architecture. The architecture of PaSCient is inspired by DeepSet Zaheer et al. (2017). The gene expression of the different cells of a given patient i is represented as a matrix $X_i \in \mathbb{R}^{M_i \times d_g}$, where M_i is the number of cells for patient i, and d_g is the number of genes measured. We first encode each cell in the sample using a learnable cell embedder function $f_\theta : \mathbb{R}^d \to d_h$, where d_h is the dimension of the cell representations. At this stage, a patient is represented as a set of vectors $\{\mathbf{z}_j : j = 1, ..., M_i\}$ of size d_h. This set can be abstracted as a matrix $Z_i \in \mathbb{R}^{M_i \times d_h}$. To create a patient-level embedding $\mathbf{e}_i$, we used a softmax-attention pooling layer:

$$\mathbf{w}_i = \text{softmax}(a_\theta(Z_i)) \tag{1}$$

$$\mathbf{e}_i = \mathbf{w}_i^T Z_i, \tag{2}$$

where $a_\theta : \mathbb{R}^{d_h} \to \mathbb{R}$ is a neural network acting on each row of Z_i independently. Lastly, the patient-level embedding is fed into a neural network classifier $h_\theta : \mathbb{R}^{d_h} \to \mathbb{R}^{d_c}$, where d_c is the number of disease classes in the pooled dataset. The final disease prediction is obtained as:

$$\hat{\mathbf{p}}_i = \text{softmax}(h_\theta(\mathbf{e}_i)), \tag{3}$$

where $\hat{\mathbf{p}}_i$ represents the predicted probabilities for each disease label. We train PaSCient end-to-end by minimizing the cross-entropy between predicted disease-state label and observed disease-state label. Different aggregation mechanisms were investigated during the development of the method. A softmax-attention layer was found to be the most effective in our ablation studies, as shown in Figure 2(b). To address the disease and tissue heterogeneity of the dataset, we introduce a dedicated sampling strategy that gives more importance to sample with low prevalence diseases and tissues. More details can be found in the Methods section.

Fine-grained importance scores. To interpret the predictions of PaSCient, we develop an approach relying on integrated gradients (IG) Sundararajan et al. (2017). This procedure starts by producing a gradient attribution for each cell-gene combination of the input sample using IG. Given the resulting matrix of attributions, we average attributions based on different dimensions, leading to different levels of interpretability. For instance, averaging the attributions over genes leads to importance scores for each individual cell, whether averaging over cells leads to importance scores for individual genes. A similar rationale can be employed to generate importance score for groups of cells (or cell types) and individual genes within a given group of cells (Figure 1(c)).

Dataset. Our dataset includes 24.3 million scRNA-seq count profiles from over 5,000 patient samples spanning 135 unique disease-state labels, across 413 studies, and 189 tissues (organs). Each patient contributed to a single sample (such that patient and samples can be used interchangeably in this text). All datasets are publicly accessible on CELLxGENE Biology et al. (2023). Cells were all profiled using droplet based scRNA-seq from 10X Genomics. The data were split into a training (60%), validation (20%), and test set (20%), ensuring that all samples from a given study are in the same split. A visual summary of our splits is described in Appendix E. The data distribution was imbalanced in terms of diseases and tissues, *e.g.* COVID-19 patients accounted for $\sim$9% of the samples, while multiple sclerosis only for $\sim$2% (Extended Data Fig. 2 (a) and (b)).

2.2 PaSCient can accurately classify disease from a patient's scRNA-seq profiles.

We train PaSCient to predict the disease label associated with each sample in the dataset and evaluate its performance in terms of weighted F1-score, a widely used metric for evaluating classification performance Abdelaal et al. (2019); Grandini et al. (2020). We compare our approach with different embedding baselines, such as a simple pseudo-bulk approach, using cell-type proportions (CTP),

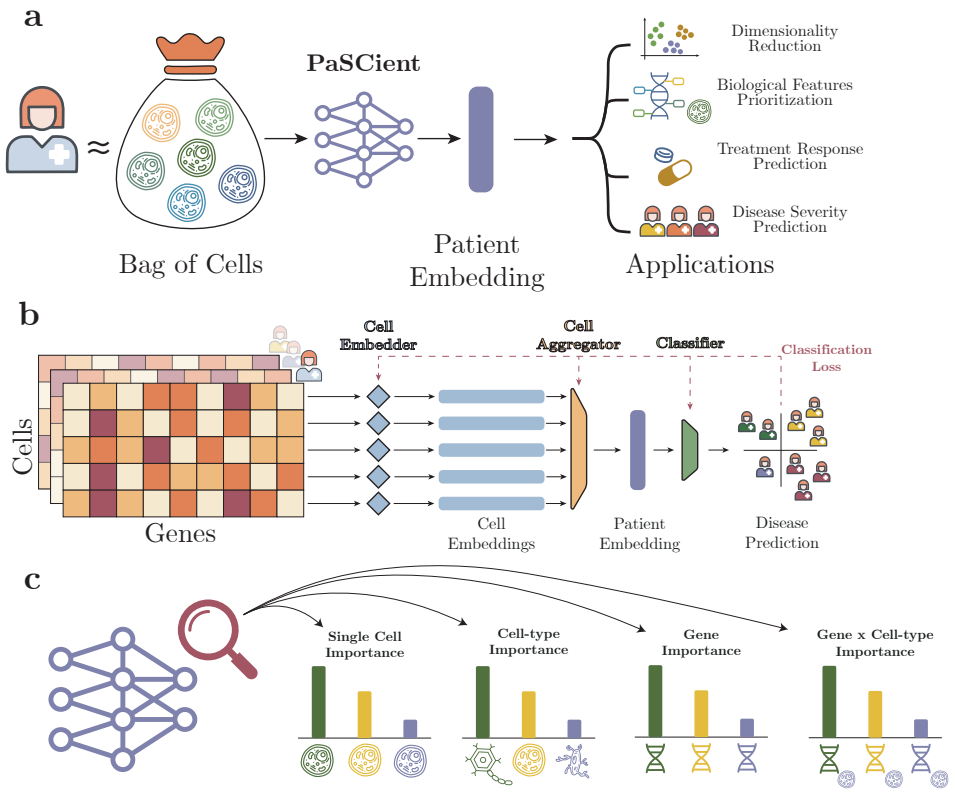


Figure 1: The landscape of PaSCient. (**a**) Model description and applications. PaSCient abstracts each patient as a bag of cells and outputs a single vector summarizing the patient's cellular context. This vector can be used for various downstream tasks, such as dimensionality reduction, visualization, biological feature prioritization, and predicting treatment response or disease severity. (**b**) Model architecture and training. Each bag of cells is represented as a gene-expression matrix, where rows correspond to individual cells, and columns represent specific genes. PaSCient first embeds each cell individually, and these cell embeddings are then summarized into a patient-level representation by a weighting the embeddings with cell-level attention. A final classifier takes this patient embedding as input to predict the disease status. The entire architecture is trained end-to-end. (**c**) Model interpretability. PaSCient enables fine-grained interpretability, generating importance scores at various levels—for individual cells, groups of cells (e.g., cell types), individual genes, or genes within specific cell groups—providing detailed insights into each patient's cellular landscape.

as well as state-of-the-art single-cell foundation models (CellPLM Wen et al. (2024) and SCimilarity Heimberg et al. (2023)). For each of these methods, we consider two classifiers to predict the label from the patient embedding: k-Nearest Neighbor Classifier (kNN) Pedregosa et al. (2011) and a multi-layer perceptron (MLP).

Remarkably, PaSCient outperforms all baselines by a significant margin (Figure 2 (a)). Notably, a simple pseudo-bulk approach outperforms more complicated foundation models in this task. Additional results on a simpler binary classification task (*i.e.,* COVID-19 vs. healthy) are given in Appendix F, including the comparison with the most recent domain-expert model ScRAT Mao et al. (2024), which performs significantly worse than PaSCient.

We investigated different aggregation mechanisms for pooling cell-level embeddings into a patient-level embedding, including mean-pooling, transformer, gated attention, linear attention, and non-linear attention mechanisms. We found that non-linear attention performed best, improving the weighted F1-score by 16.6% compared to a mean-pooling mechanism (Figure 2(b)). The transformer approach, although more expressive, results in poor performance, probably due to a larger than necessary number of parameters for this task.

To account for the class imbalance in the data, we investigated different resampling mechanisms. We studied the impact of resampling both per disease-class and per tissue-class (Methods). Oversampling the training set for both disease and tissue resulted in a significant improvement compared to baseline (Figure 2(b)). Model training and hyper-parameter tuning details are given in Appendix G.

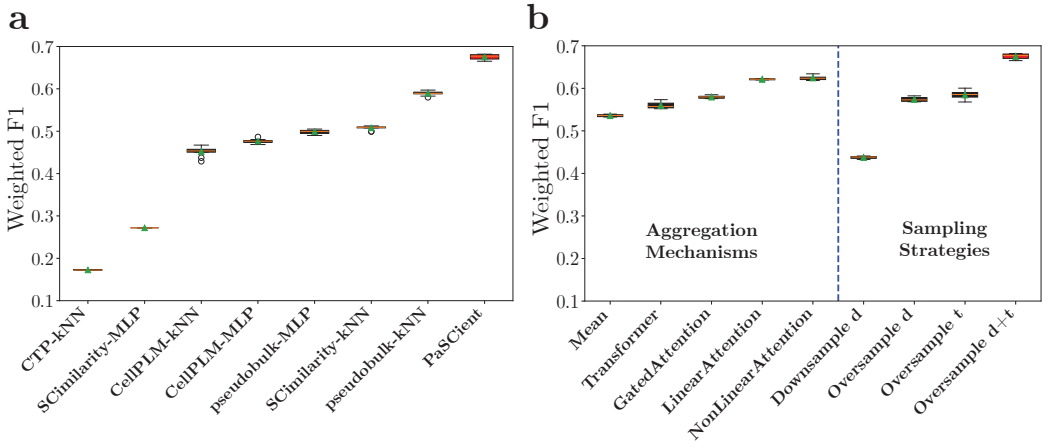


Figure 2: Benchmarking the performance of PaSCient on multi-disease classification. (**a**) Weighted F1-score results. Performance comparison between PaSCient and relevant baseline models, with standard deviations calculated from experiments using different seeds. PaSCientemploys non-linear attention aggregation combined with oversampling based on disease and tissue. (**b**) Ablation studies. Analysis of different training configurations for PaSCient, including various cell-level aggregation methods (without resampling) and sampling strategies to address label imbalance. The best performance was achieved using non-linear attention aggregation with oversampling based on both disease and tissue labels (Oversample d+t).

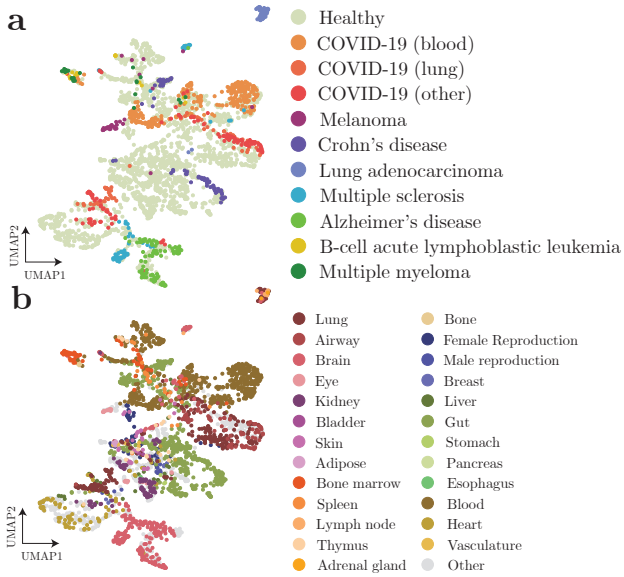


Figure 3: Patient embeddings using PaSCient organize by both tissue and disease. Uniform manifold approximation and projection (UMAP) of patient embeddings colored by each of 8 most common disease labels (**a**) or by tissue (**b**). We only visualize the samples whose disease-state labels exist in all the splits.

The patient embedding space learned by PaSCient is organized by disease state (Figure 3(a)) and by tissue (Figure 3(b)). Notably, COVID-19 patients partition into two clusters, corresponding to blood and lung tissue samples. Additional analyses of the patient embedding space, aggregated per disease, are given in Appendices H and I.

2.3 PaSCient prioritizes gene and cell-type roles in disease prediction.

We use our importance score methodology (described in Section 2.1 and in the Methods section) to enable a fine-grained analysis of the individual cells and genes that contribute most to a disease of interest. As a proof of concept, we focus our analysis on COVID-19 prediction and select a cohort of patients with a COVID-19 disease label.

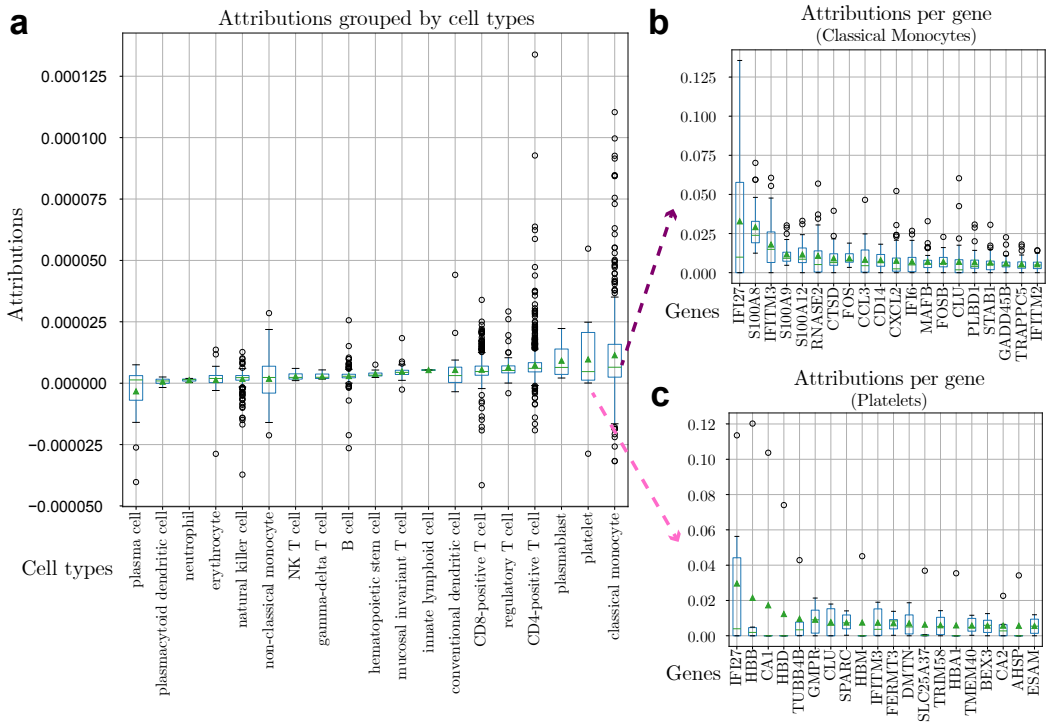


Figure 4: Prioritizing cell types and cell-type-specific genes for COVID-19 by integrated gradients (IG) analysis. (**a**) Attributions averaged over cells and genes for each cell type (each point is a patient). Cell types are ranked by their mean attribution, with classical monocytes and platelets identified as the most predictive for COVID-19 diagnosis. (**b**) Attributions aggregated over classical monocytes (each point is a patient). Genes are ranked by mean attribution, with the green line indicating the median value and the green triangle denoting the mean value. (**c**) Attributions aggregated over platelets (each point is a patient). Genes are ranked by mean attribution, with the green line indicating the median value and the green triangle denoting the mean value.

We first compute cell type level attributions to uncover what cell types were contributing most to the COVID-19 label for each patient (Figure 4(a)). The highest average attributions (computed over all patients) are found for classical monocytes and platelets, suggesting the importance of these cell types in COVID-19. Notably, these cell types have been identified in the literature as playing a key-role in the disease pathogenesis Junqueira et al. (2022); Wool & Miller (2021).

Remarkably, our fine-grained importance methodology enables further exploration within cell types of interest. We investigate what genes were most impacting COVID-19 prediction for each of these cell types specifically. For each patient, we compute the importance of gene in monocytes (Figure 4(b)) and in platelets (Figure 4(b)). This procedure identifies the specific importance of genes in a given cell type. Ranking genes by average importance reveals that S100A8, IFITM3, and IFI27 are the most pertinent genes in monocytes for COVID-19. IFI27, HBB, and CA1 are found to be most important in platelets. These genes are associated with COVID-19 severity or treatment Mellett & Khader (2022); Xu et al. (2022); Shojaei et al. (2023); Zhang et al. (2022); Deniz et al. (2021).

We validate the set of important genes uncovered by PaSCient by measuring the overlap with the set of differentially expressed genes from ToppCell Jin et al. (2021). A Fisher's exact test indicates strong overlap for both classical monocytes (p-value=2.1e-22) and platelets (p-value=2.5e-20). A similar analysis for other diseases is presented in Appendix J. These analyses show that we can capture and prioritize disease-specific genes and cell types at different resolutions.

2.4 PaSCient recovers disease severity of individual patients.

To investigate the patient representations learnt by our method, we collect four scRNA-seq datasets from COVID-19 patients where a severity label is available (mild or severe) Lemsara et al. (2022); Schulte-Schrepping et al. (2020); Wilk et al. (2021); Lee et al. (2020), and that were not included during training. Visualizing the patient representations generated by our model, we find that the landscape is primarily organized by disease severity and not by study (Figure 5(a)). Conversely, a

principal components analysis (PCA) representation of pseudo-bulk data is organized primarily by study rather than severity, highlighting batch effects.

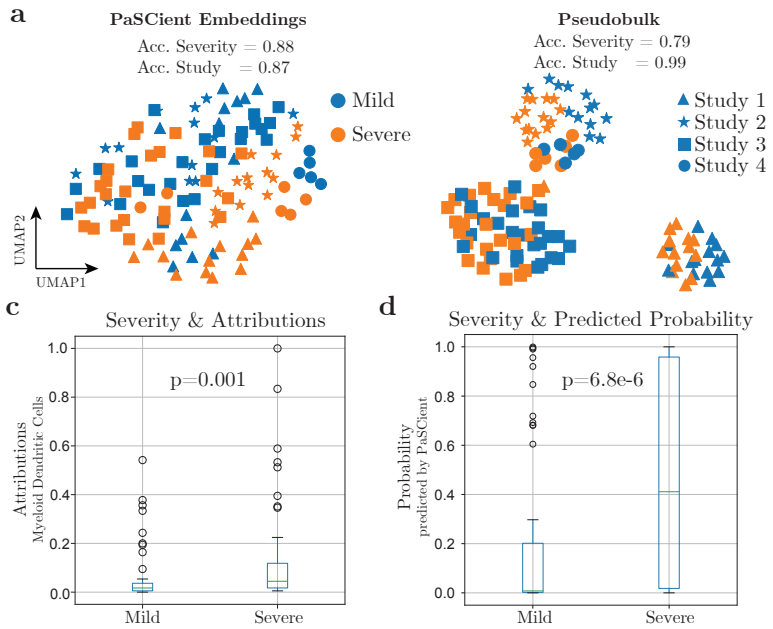


Figure 5: PaSCient captures disease severity in COVID-19 patients. (**a**) Patient embeddings generated by PaSCient and PCA on pseudo-bulk data, colored by disease severity. PaSCient organizes patient representations based on disease severity, whereas the pseudo-bulk embeddings are influenced by study-specific effects. The accuracy of a k-nearest-neighbor (kNN) classifier is reported for both disease severity and study labels to quantitatively assess embedding quality. Higher accuracy suggests the embedding is more organized according to that specific variable. (**b**) Magnitude of integrated gradients attributions averaged across all myeloid dendritic cells for each sample, grouped by disease severity. P-values are Bonferroni-corrected. (**c**) Probability of COVID-19 diagnosis predicted by PaSCient for each sample, stratified by disease severity.

Moreover, the importance scores given by our model to different cell types correlates with disease severity, with significant associations (corrected $p < 0.01$) for NK cells, B cells, myeloid dendritic cells, and MAIT cells. Indeed, there is a significant difference in the magnitude of the integrated gradients attributions of the model, averaged over all myeloid dendritic cells in each patient sample, between mild and severe patient groups (Figure 5(b), Bonferroni-corrected p-value=0.001, rank sum test). Similarly, there is a significant association between the disease severity and the magnitude of the probability of COVID-19 diagnosis predicted by PaSCient (Figure 5(c)). Together, these results show that PaSCient can implicitly represent the disease severity of each patient. Associations between severity and other cell types are given in Appendix K. A case study for predicting drug response is presented in Appendix L.

3 Discussion

Here, we introduced a new model, PaSCient, that generates patient-level embeddings given a single-cell RNA-seq context, leveraging thousands of samples. PaSCient builds upon recent single-cell foundation models Cui et al. (2024); Heimberg et al. (2023); Hao et al. (2024), and multi-cellular representations models He et al. (2021); Mao et al. (2024); Xiong et al. (2023) but differs in key aspects. First, PaSCient builds upon the large scale training of single-cells foundation models but extends the approach to multi-cellular representations. While single-cell representations can be pooled into a patient-level representation (*e.g.,* via average-pooling), our experiments showed that this resulted in sub-optimal performance. Our approach is indeed more expressive as it learns a dedicated aggregation mechanism that better reflects the underlying biological processes. Second, PaSCient extends previous works on multi-cellular representations by going beyond binary classifications and by leveraging hundreds of single-cell expression studies.

Providing biologically informed patient-level representations presents several advantages for biological and clinical research. Such representations enable a patient-specific understanding of disease

mechanisms and can improve patient segmentation, thereby contributing to more targeted therapies. We demonstrated the potential of PaSCient in patient segmentation by showing that the learnt embeddings implicitly encoded clinical information such as disease severity. From a target discovery perspective, we highlighted the fine-grained resolution of our importance scores. We showed that our model could be used to prioritize individual cells and genes, but also groups of cells (such as cell types) and cell-type-specific genes, underlining a promising knowledge discovery toolkit.

Our work represents an important step toward patient-level representations contextualized by single-cell expression. While our datasets included millions of cells, the increasing scale of available single-cell repositories suggests further iterations of this class of models will lead to better representations.

4 Methods

Notations. We define the aggregated dataset includes N patient samples: $\mathcal{D} = \{s_1, s_2, ..., s_N\}$, where s_i represents the i_{th} patient sample. Each patient includes M_i cells (where M_i varies per patient): $s_i = \{c_1, c_2, ...c_{M_i}\}_i$, where c_j represents the j_{th} cell in s_i. Lastly, each cell c_j is a vector whose features are gene expression counts with dimension $d_g = 28,231$. Each patient can then be represented as a matrix $X_i \in \mathbb{R}^{M_i \times d_g}$. Patient-level metadata is also available such as disease label y_i and tissue label t_i.

Model architecture. PaSCient combines a cell encoder $f_\theta(\cdot)$, an aggregator $h_\theta(\cdot)$, and a classifier $g_\theta(\cdot)$, all implemented by neural networks. At a high level, the cell encoder produces an embedding for each cell in a patient sample, the aggregator combines the cell embeddings into a patient embedding, and the classifier predicts the disease label based on the patient embedding.

The cell encoder is a linear layer. The classifier is a multi-layer perceptron (MLP) with a final softmax activation. We write the output of the cell encoder as $z_i = f_\theta(c_i)$, the output of the aggregator as $\mathbf{e}_i = g_\theta(\mathbf{z}_i)$ with $\mathbf{z} = [\mathbf{z}_1, \mathbf{z}_2, ..., \mathbf{z}_{M_i}]_i)$, and the output of the classifier as $\hat{p}_i = h_\theta(\mathbf{e}_i)$. The model is trained by minimizing the cross-entropy between $\hat{\mathbf{p}}_i$ and y_i. A graphical depiction of the model architecture is given in Figure 1.

Aggregators. We considered multiple different aggregators. Most aggregators have the form of a weighted sum: $\mathbf{e}_i = \sum_{j=1}^{M_i} w_j \mathbf{z}_j$. Aggregators differ by the way the weights $\mathbf{w} = [w_1, ..., w_{M_i}]$ are computed. The mean aggregator uses $w_j = \frac{1}{M_i}$; the linear attention aggregator uses $\mathbf{w} = Softmax(\mathbf{z})$; the non-linear attention uses $Softmax(a_\theta(\mathbf{z}))$ with a_θ a learnable neural network that operates on each $\mathbf{z}_j$ independently; and the gated-attention uses $\mathbf{w} = Softmax(U_\theta(\mathbf{z}) \odot Sigmoid(V_\theta(\mathbf{z})))$ with two learnable neural networks u_θ and v_θ. The transformer aggregator differs in its architecture as it updates the embeddings of each cell according to the entire sample and sums the resulting embeddings.

Resampling strategies. We used the following resampling strategies for addressing the disease and tissue imbalances in the dataset: (1) Downsampling disease: subsampling the most frequent disease classes such as to balance the disease label overall; (2) Oversampling disease: oversampling the least frequent disease classes such as to balance the disease label overall; (3) Oversampling tissue: oversampling the least frequent tissue classes such as to balance the tissue label overall; (4) Oversampling disease and tissue: oversampling the least frequent tissue and disease classes such as to balance both tissue and disease labels overall.

Model explainability. We used the integrated gradients method on the input matrix X_i Sundararajan et al. (2017). Computing the integrated gradients on this input results in an attribution matrix $R_i \in \mathbb{R}^{M_i \times d_g}$ with the same dimensions as the input matrix. The attribution of a given gene was obtained by averaging R_i across all cells. The attribution of a given cell was obtained by averaging over all genes. Any other combination follows from generalizing this procedure.

Disease classification metrics. We evaluated classification performance using the weighted F1-score. F1-score is robust to class imbalance and reflects both precision and recall across all classes.

Each experiment was repeated 10 times using different seeds leading to different cells being sampled for each patient. This repetition allowed computing an empirical standard deviation on the results.

Dataset pre-processing. All datasets were profiled by droplet based scRNA-Seq from 10X Genomics. We removed cell profiles with no gene expression levels and normalized all remaining profiles to the corrected sequencing depth, followed by a $log(x + 1)$ transformation.

Reproducibility and Data The sources of datasets used for training/validating/testing as well as downstream applications can be found in the Supplementary File 1. Our collected descriptions for diseases and tissues can be found in Supplementary File 2. The genes from ToppCell are listed in Supplementary File 3. The running time of our method for different tasks is included in Supplementary File 4.

References

Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. *Nature medicine*, 27(11):1876–1884, 2021.

Tamim Abdelaal, Lieke Michielsen, Davy Cats, Dylan Hoogduin, Hailiang Mei, Marcel JT Reinders, and Ahmed Mahfouz. A comparison of automatic cell identification methods for single-cell rna sequencing data. *Genome biology*, 20:1–19, 2019.

Jeremiah Olawumi Arowoogun, Oloruntoba Babawarun, Rawlings Chidi, Adekunle Oyeyemi Adeniyi, and Chioma Anthonia Okolo. A comprehensive review of data analytics in healthcare management: Leveraging big data for decision-making. *World Journal of Advanced Research and Reviews*, 21(2):1810–1821, 2024.

Tanya Barrett, Stephen E Wilhite, Pierre Ledoux, Carlos Evangelista, Irene F Kim, Maxim Tomashevsky, Kimberly A Marshall, Katherine H Phillippy, Patti M Sherman, Michelle Holko, et al. Ncbi geo: archive for functional genomics data sets—update. *Nucleic acids research*, 41(D1): D991–D995, 2012.

CZI Single-Cell Biology, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, et al. Cz cellxgene discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *BioRxiv*, pp. 2023–10, 2023.

Pierre Boyeau, Justin Hong, Adam Gayoso, Martin Kim, José L McFaline-Figueroa, Michael I Jordan, Elham Azizi, Can Ergen, and Nir Yosef. Deep generative modeling of sample-level heterogeneity in single-cell genomics. *BioRxiv*, pp. 2022–10, 2022.

Yan-Mei Chen, Yuanting Zheng, Ying Yu, Yunzhi Wang, Qingxia Huang, Feng Qian, Lei Sun, Zhi-Gang Song, Ziyin Chen, Jinwen Feng, et al. Blood molecular markers associated with covid-19 immunopathology and multi-organ damage. *The EMBO journal*, 39(24):e105896, 2020.

Shih-Kai Chu, Shilin Zhao, Yu Shyr, and Qi Liu. Comprehensive evaluation of noise reduction methods for single-cell rna sequencing data. *Briefings in bioinformatics*, 23(2):bbab565, 2022.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, Nan Duan, and Bo Wang. scgpt: toward building a foundation model for single-cell multi-omics using generative ai. *Nature Methods*, pp. 1–11, 2024.

Sabyasachi Dash, Sushil Kumar Shakyawar, Mohit Sharma, and Sandeep Kaushik. Big data in healthcare: management, analysis and future prospects. *Journal of big data*, 6(1):1–25, 2019.

Saloni Dattani, David M Howard, Cathryn M Lewis, and Pak C Sham. Clarifying the causes of consistent and inconsistent findings in genetics. *Genetic epidemiology*, 46(7):372–389, 2022.

Secil Deniz, Tugba Kevser Uysal, Clemente Capasso, Claudiu T Supuran, and Ozen Ozensoy Guler. Is carbonic anhydrase inhibition useful as a complementary therapy of covid-19 infection? *Journal of Enzyme Inhibition and Medicinal Chemistry*, 36(1):1230–1235, 2021.

Maria Teresa Ferretti, Maria Florencia Iulita, Enrica Cavedo, Patrizia Andrea Chiesa, Annemarie Schumacher Dimech, Antonella Santuccione Chadha, Francesca Baracchi, Hélène Girouard, Sabina Misoch, Ezio Giacobini, et al. Sex differences in alzheimer disease—the gateway to precision medicine. *Nature Reviews Neurology*, 14(8):457–469, 2018.

Margherita Grandini, Enrico Bagli, and Giorgio Visani. Metrics for multi-class classification: an overview. *arXiv preprint arXiv:2008.05756*, 2020.

Minsheng Hao, Jing Gong, Xin Zeng, Chiming Liu, Yucheng Guo, Xingyi Cheng, Taifeng Wang, Jianzhu Ma, Xuegong Zhang, and Le Song. Large-scale foundation model on single-cell transcriptomics. *Nature Methods*, pp. 1–11, 2024.

Bryan He, Matthew Thomson, Meena Subramaniam, Richard Perez, Chun Jimmie Ye, and James Zou. Cloudpred: Predicting patient phenotypes from single-cell rna-seq. pp. 337–348, 2021.

Graham Heimberg, Tony Kuo, Daryle DePianto, Tobias Heigl, Nathaniel Diamant, Omar Salem, Gabriele Scalia, Tommaso Biancalani, Shannon Turley, Jason Rock, et al. Scalable querying of human cell atlases via a foundational model reveals commonalities across fibrosis-associated macrophages. *bioRxiv*, pp. 2023–07, 2023.

Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

Huixiao Hong, Lei Xu, Zhenqiang Su, Jie Liu, Weigong Ge, Jie Shen, Hong Fang, Roger Perkins, Leming Shi, and Weida Tong. Pitfall of genome-wide association studies: Sources of inconsistency in genotypes and their effects. 2012.

John PA Ioannidis. Non-replication and inconsistency in the genome-wide association setting. *Human heredity*, 64(4):203–213, 2007.

Philipp Janssen, Zane Kliesmete, Beate Vieth, Xian Adiconis, Sean Simmons, Jamie Marshall, Cristin McCabe, Holger Heyn, Joshua Z Levin, Wolfgang Enard, et al. The effect of background noise and its removal on the analysis of single-cell expression data. *Genome Biology*, 24(1):140, 2023.

Kang Jin, Eric E Bardes, Alexis Mitelpunkt, Jake Y Wang, Surbhi Bhatnagar, Soma Sengupta, Daniel P Krummel, Marc E Rothenberg, and Bruce J Aronow. A web portal and workbench for biological dissection of single cell covid-19 host responses. *bioRxiv*, pp. 2021–06, 2021.

Aashi Jindal, Prashant Gupta, Jayadeva, and Debarka Sengupta. Discovery of rare cells from voluminous single cell expression data. *Nature communications*, 9(1):4719, 2018.

Caroline Junqueira, Ângela Crespo, Shahin Ranjbar, Luna B De Lacerda, Mercedes Lewandrowski, Jacob Ingber, Blair Parry, Sagi Ravid, Sarah Clark, Marie Rose Schrimpf, et al. Fcγr-mediated sars-cov-2 infection of monocytes activates inflammation. *Nature*, 606(7914):576–584, 2022.

Minoru Kanehisa. Toward understanding the origin and evolution of cellular organisms. *Protein Science*, 28(11):1947–1951, 2019.

Minoru Kanehisa and Susumu Goto. Kegg: kyoto encyclopedia of genes and genomes. *Nucleic acids research*, 28(1):27–30, 2000.

Minoru Kanehisa, Miho Furumichi, Yoko Sato, Masayuki Kawashima, and Mari Ishiguro-Watanabe. Kegg for taxonomy-based analysis of pathways and genomes. *Nucleic acids research*, 51(D1): D587–D592, 2023.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Andrew L Kau and Phillip E Korenblat. Anti-interleukin 4 and 13 for asthma treatment in the era of endotypes. *Current opinion in allergy and clinical immunology*, 14(6):570–575, 2014.

Jeong Seok Lee, Seongwan Park, Hye Won Jeong, Jin Young Ahn, Seong Jin Choi, Hoyoung Lee, Baekgyu Choi, Su Kyung Nam, Moa Sa, Ji-Soo Kwon, et al. Immunophenotyping of covid-19 and influenza highlights the role of type i interferons in development of severe covid-19. *Science immunology*, 5(49):eabd1554, 2020.

Jeffrey T Leek, Robert B Scharpf, Héctor Corrada Bravo, David Simcha, Benjamin Langmead, W Evan Johnson, Donald Geman, Keith Baggerly, and Rafael A Irizarry. Tackling the widespread and critical impact of batch effects in high-throughput data. *Nature Reviews Genetics*, 11(10): 733–739, 2010.

Amina Lemsara, Adrian Chan, Dominik Wolff, Michael Marschollek, Yang Li, and Christoph Dieterich. Robust machine learning predicts covid-19 disease severity based on single-cell rna-seq from multiple hospitals. *medRxiv*, pp. 2022–10, 2022.

Zhexiao Lin and Wei Sun. Supervised deep learning with gene annotation for cell classification. *bioRxiv*, pp. 2024–07, 2024.

Anastasia Litinetskaya, Maiia Shulman, Soroor Hediyeh-zadeh, Amir Ali Moinfar, Fabiola Curion, Artur Szalata, Alireza Omidi, Mohammad Lotfollahi, and Fabian J Theis. Multimodal weakly supervised learning to identify disease-specific changes in single-cell atlases. *bioRxiv*, pp. 2024–07, 2024.

Tianyu Liu, Wenxin Long, Zhiyuan Cao, Yuge Wang, Chuan Hua He, Le Zhang, Stephen M Strittmatter, and Hongyu Zhao. Cosgenegate selects multi-functional and credible biomarkers for single-cell analysis. *bioRxiv*, pp. 2024–05, 2024.

Yuzhen Mao, Yen-Yi Lin, Nelson KY Wong, Stanislav Volik, Funda Sar, Colin Collins, and Martin Ester. Phenotype prediction from single-cell rna-seq data using attention-based neural networks. *Bioinformatics*, pp. btae067, 2024.

Vivien Marx. The big challenges of big data. *Nature*, 498(7453):255–260, 2013.

Leah Mellett and Shabaana A Khader. S100a8/a9 in covid-19 pathogenesis: Impact on clinical outcomes. *Cytokine & Growth Factor Reviews*, 63:90–97, 2022.

Jonathan Mitchel, M Grace Gordon, Richard K Perez, Evan Biederstedt, Raymund Bueno, Chun Jimmie Ye, and Peter V Kharchenko. Coordinated, multicellular patterns of transcriptional variation that stratify patient cohorts are revealed by tensor decomposition. *Nature Biotechnology*, pp. 1–10, 2024.

Kevin Musgrave, Serge J. Belongie, and Ser-Nam Lim. Pytorch metric learning. *ArXiv*, abs/2008.09164, 2020.

Ziad Obermeyer and Ezekiel J Emanuel. Predicting the future—big data, machine learning, and clinical medicine. *New England Journal of Medicine*, 375(13):1216–1219, 2016.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun

Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. Gpt-4 technical report, 2024. URL `https://arxiv.org/abs/2303.08774`.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

Jeffrey M Pullin and Davis J McCarthy. A comparison of marker gene selection methods for single-cell rna sequencing data. *Genome Biology*, 25(1):56, 2024.

Aviv Regev, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, Peter Campbell, Piero Carninci, Menna Clatworthy, et al. The human cell atlas. *elife*, 6:e27041, 2017.

Jennifer E Rood, Aidan Maartens, Anna Hupalowska, Sarah A Teichmann, and Aviv Regev. Impact of the human cell atlas on medicine. *Nature medicine*, 28(12):2486–2496, 2022.

Moshe Sade-Feldman, Keren Yizhak, Stacey L Bjorgaard, John P Ray, Carl G de Boer, Russell W Jenkins, David J Lieb, Jonathan H Chen, Dennie T Frederick, Michal Barzily-Rokni, et al. Defining t cell states associated with response to checkpoint immunotherapy in melanoma. *Cell*, 175 (4):998–1013, 2018.

Eric W Sayers, Jeff Beck, Evan E Bolton, J Rodney Brister, Jessica Chan, Donald C Comeau, Ryan Connor, Michael DiCuccio, Catherine M Farrell, Michael Feldgarden, et al. Database resources of the national center for biotechnology information. *Nucleic Acids Research*, 52(D1):D33, 2024.

Daniel J Schaid, Wenan Chen, and Nicholas B Larson. From genome-wide associations to candidate causal variants by statistical fine-mapping. *Nature Reviews Genetics*, 19(8):491–504, 2018.

Jonas Schulte-Schrepping, Nico Reusch, Daniela Paclik, Kevin Baßler, Stephan Schlickeiser, Bowen Zhang, Benjamin Krämer, Tobias Krammer, Sophia Brumhard, Lorenzo Bonaguro, et al. Severe covid-19 is marked by a dysregulated myeloid cell compartment. *Cell*, 182(6):1419–1440, 2020.

Maryam Shojaei, Amir Shamshirian, James Monkman, Laura Grice, Minh Tran, Chin Wee Tan, Siok Min Teo, Gustavo Rodrigues Rossi, Timothy R McCulloch, Marek Nalos, et al. Ifi27 transcription is an early predictor for covid-19 outcomes, a multi-cohort observational study. *Frontiers in Immunology*, 13:1060438, 2023.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017.

Mario L Suvà, Esther Rheinbay, Shawn M Gillespie, Anoop P Patel, Hiroaki Wakimoto, Samuel D Rabkin, Nicolo Riggi, Andrew S Chi, Daniel P Cahill, Brian V Nahed, et al. Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. *Cell*, 157 (3):580–594, 2014.

Aliasghar Tarkhan, Trung Kien Nguyen, Noah Simon, and Jian Dai. Survival prediction via deep attention-based multiple-instance learning networks with instance sampling, 2023.

Hanchen Wang, Jure Leskovec, and Aviv Regev. Metric mirages in cell embeddings. *bioRxiv*, pp. 2024–04, 2024.

Hongzhi Wen, Wenzhuo Tang, Xinnan Dai, Jiayuan Ding, Wei Jin, Yuying Xie, and Jiliang Tang. CellPLM: Pre-training of cell language model beyond single cells, 2024. URL `https://openreview.net/forum?id=BKXvPDekud`.

Aaron J Wilk, Madeline J Lee, Bei Wei, Benjamin Parks, Ruoxi Pi, Giovanny J Martínez-Colón, Thanmayi Ranganath, Nancy Q Zhao, Shalina Taylor, Winston Becker, et al. Multi-omic profiling reveals widespread dysregulation of innate immunity and hematopoiesis in covid-19. *Journal of Experimental Medicine*, 218(8):e20210582, 2021.

Geoffrey D Wool and Jonathan L Miller. The impact of covid-19 disease on platelets and coagulation. *Pathobiology*, 88(1):15–27, 2021.

Guangzhi Xiong, Stefan Bekiranov, and Aidong Zhang. Protocell4p: an explainable prototype-based neural network for patient classification using single-cell rna-seq. *Bioinformatics*, 39(8):btad493, 2023.

Fengwen Xu, Geng Wang, Fei Zhao, Yu Huang, Zhangling Fan, Shan Mei, Yu Xie, Liang Wei, Yamei Hu, Conghui Wang, et al. Ifitm3 inhibits sars-cov-2 infection and is associated with covid-19 susceptibility. *Viruses*, 14(11):2553, 2022.

Fuzhao Xue, Yao Fu, Wangchunshu Zhou, Zangwei Zheng, and Yang You. To repeat or not to repeat: Insights from scaling llm under token-crisis. *Advances in Neural Information Processing Systems*, 36, 2024.

Kathryn E Yost, Ansuman T Satpathy, Daniel K Wells, Yanyan Qi, Chunlin Wang, Robin Kageyama, Katherine L McNamara, Jeffrey M Granja, Kavita Y Sarin, Ryanne A Brown, et al. Clonal replacement of tumor-specific t cells following pd-1 blockade. *Nature medicine*, 25(8):1251–1259, 2019.

Manzil Zaheer, Satwik Kottur, Siamak Ravanbakhsh, Barnabas Poczos, Russ R Salakhutdinov, and Alexander J Smola. Deep sets. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL `https://proceedings.neurips.cc/paper_files/paper/2017/file/f22e4747da1aa27e363d86d40ff442fe-Paper.pdf`.

Zili Zhang, Fanjie Lin, Fei Liu, Qiongqiong Li, Yuanyuan Li, Zhanbei Zhu, Hua Guo, Lidong Liu, Xiaoqing Liu, Wei Liu, et al. Proteomic profiling reveals a distinctive molecular signature for critically ill covid-19 patients compared with asthma and chronic obstructive pulmonary disease. *International Journal of Infectious Diseases*, 116:258–267, 2022.

Shuncang Zhu, Tao Ge, Junjie Hu, Gening Jiang, and Peng Zhang. Prognostic value of surgical intervention in advanced lung adenocarcinoma: a population-based study. *Journal of Thoracic Disease*, 13(10):5942, 2021.

A Acknowledgements

We thank Jenna Collier, Max Gold, Velina Kozareva, Gokcen Eraslan, Chainglin Wan, David Garfield, and Runming Wei for their suggestions that strengthened the quality of experiments and manuscript.

B Contributions

TL conceived of the method with input from GH and AR. TL proposed the method and finished experiments with the help of TK, EDB, GS, ND, HW, MH, AM, and HCB. TL and EDB wrote the manuscript with input from GH, GS, AM, and AR. GS, AR, and GH jointly supervised this work.

C Competing interests

All authors are employees of Genentech or Roche. A.R. is a co-founder and equity holder of Celsius Therapeutics, an equity holder in Immunitas, and until July 31, 2020 was an S.A.B. member of Thermo Fisher Scientific, Syros Pharmaceuticals, Neogene Therapeutics and Asimov.

D Reproducibility, data, and code availability

The sources of datasets used for training/validating/testing as well as downstream applications can be found in the Supplementary File 1. Our collected descriptions for diseases and tissues can be found in Supplementary File 2. The genes from ToppCell are listed in Supplementary File 3.

We used a server with eight NVIDIA A100 GPUs and 300 GB maximal RAM to conduct all the experiments. The minimal requirement for training/inference based on our model is one A100 GPU, 80 GB. The running time of our method for different tasks is included in Supplementary File 4.

All the codes used in model training and downstream applications can be found in `https://github.com/edebrouwer/pascient`.

E Dataset

In Extended Data Figure 1, we present a graphical overview of the dataset used for training our model. Extended Data Figure 1 shows the proportions of diseases and tissues in the dataset.

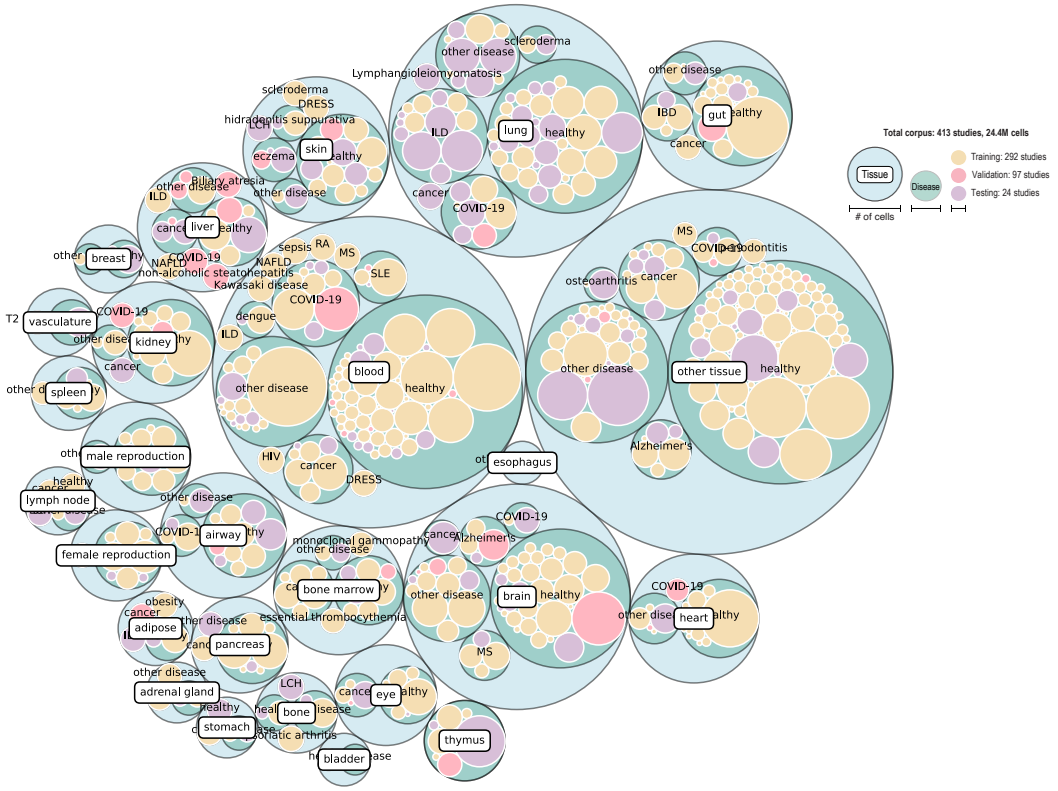
F Binary disease classification with PaSCient

We evaluated the performance PaSCient to classify patients with health and COVID-19 labels and benchmark it with other methods for performing binary disease-health classification. The results are given in Extended Data Figure 3.

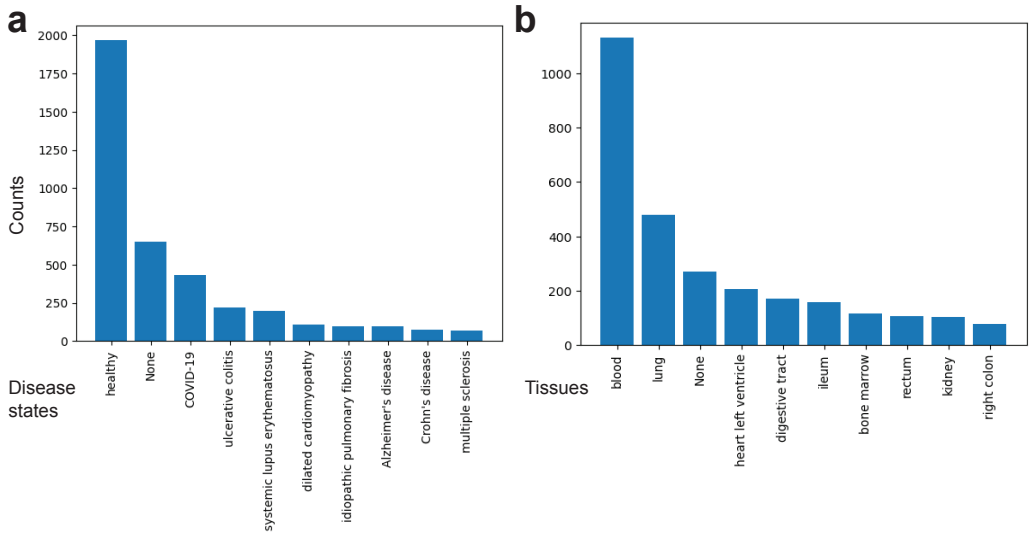
We found that PaSCient outperformed all baselines by a significant margin. We also found that linear attention aggregation performed best than other mechanisms.

F.1 Including cell-type information in the prediction

We further investigated whether introducing cell-type information can help predicting the disease label more accurately. We designed a new baseline by using the cell-type proportions to represent each patient and classified the binary conditions based on a kNN classifier (CTP-kNN). We also introduce a modified loss function for PaSCient by using a contrastive learning approach Musgrave et al. (2020) to bring cells from the same type closer together in embedding space, while pushing apart cells of different types (PaSCient-CT). Such an approach can reduce the potential batch effect existing in the training data Heimberg et al. (2023). The results are given in Extended Data Figure 3(c). We observed that PaSCient achieved the best performance, followed by our modified loss function (PaSCient-CT) and the cell type proportions baselines.



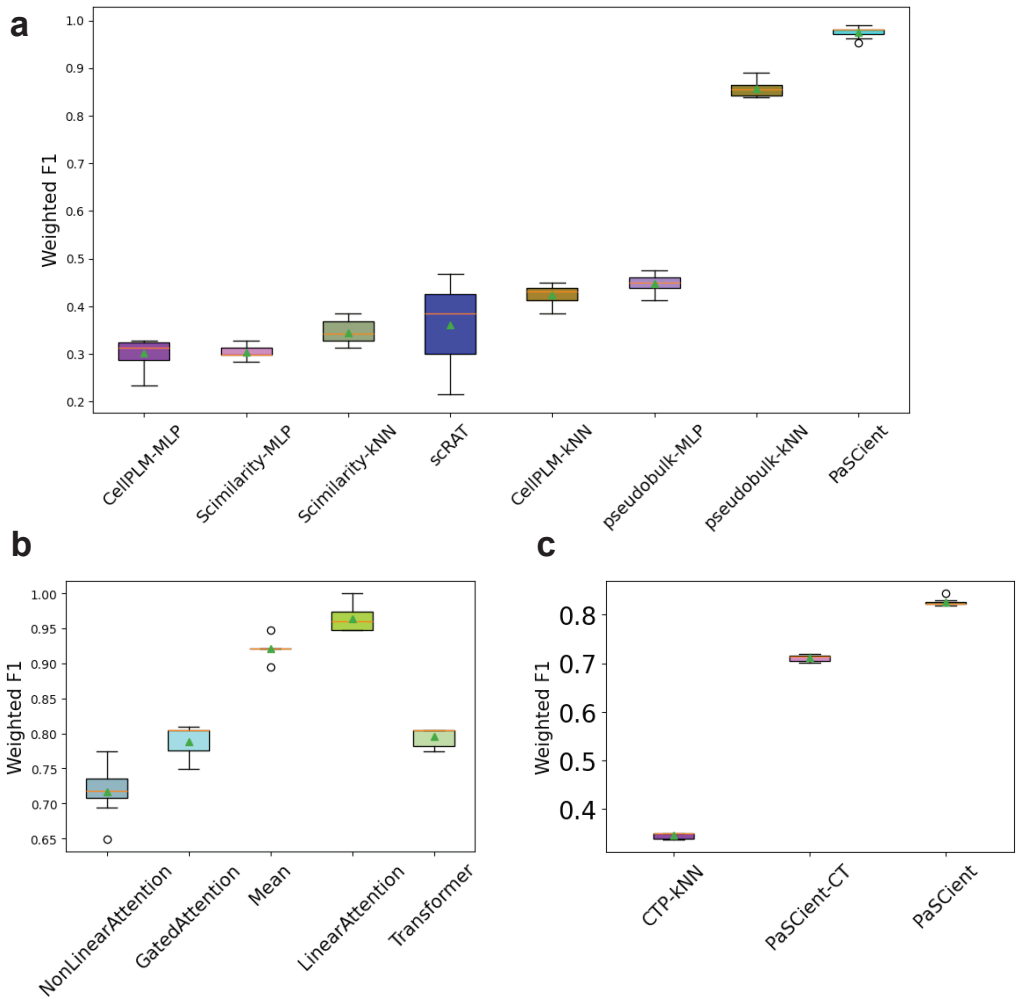
Extended Data Fig. 1: Overview of our training datasets colored by tissues and diseases. The length of bubbles represent number of cells included in the given tissue/diseases.



Extended Data Fig. 2: Statistics of disease states and tissues in our collected datasets.

G Model training details

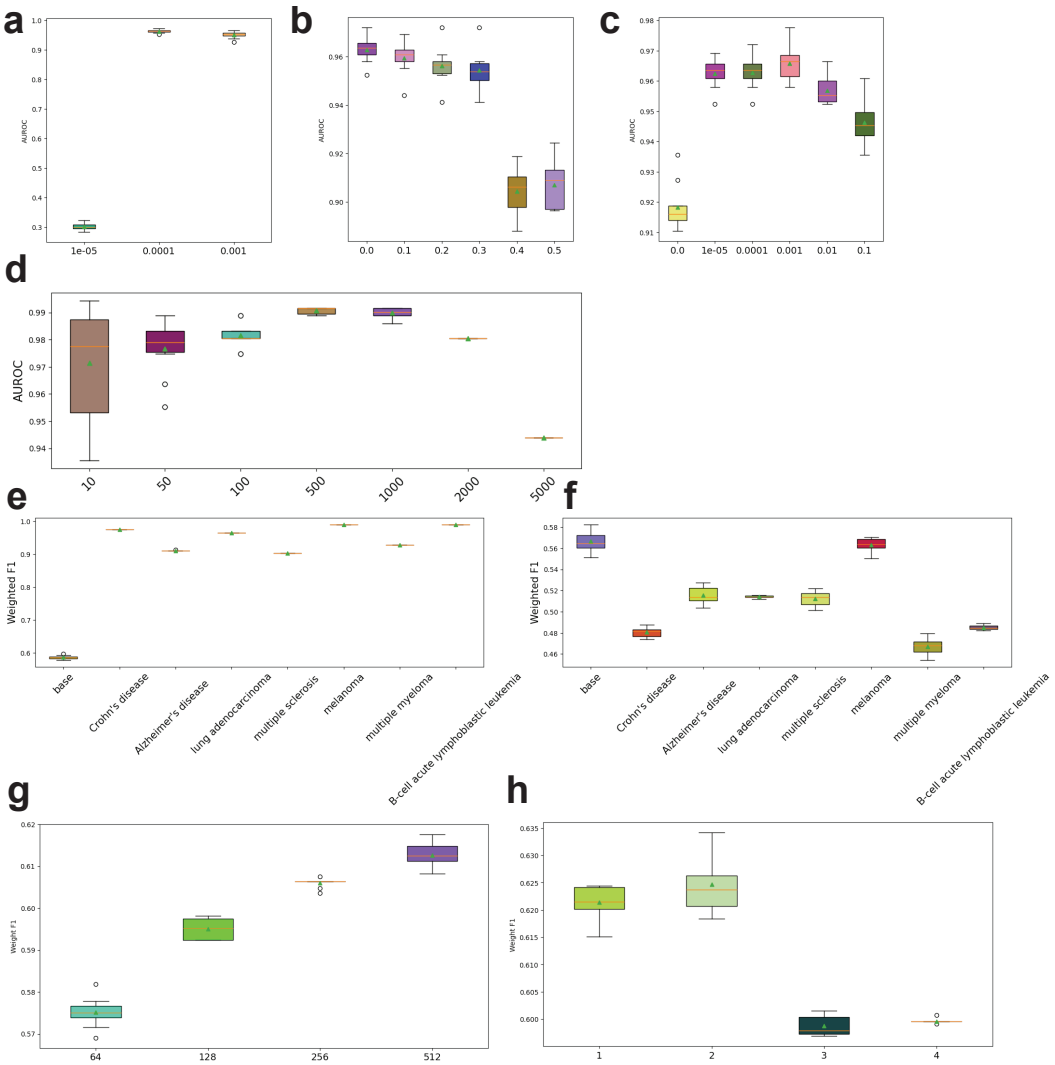
During the training process of PaSCient, we explored different factors that could affect the training process, including hyper-parameters, size of sampled cells, composition of diseases, and scaling law Hernandez et al. (2021); Kaplan et al. (2020). The sensitivity analyses focusing on these factors could help us understand the difficulties of patient modeling in a broader vision. We found that



Extended Data Fig. 3: Benchmarking results for binary classification. (**a**): Comparisons between PaSCient and the rest of baselines for classifying COVID-19 versus healthy condition. (**b**): Ablation study for different aggregation mechanisms. (**c**) Impact of cell type labels on performance. (PaSCient-CL) uses a modified contrastive training approach that uses cell type labels. CTP-kNN is kNN disease classifier based on cell types proportions for each sample.

learning rate closing to 1e-4 (illustrated in Extended Data Figure 4 (a)) could reduce the negative effects brought by over-fitting. Meanwhile, a small number of epochs (< 40 for binary classification and < 5 for multi-label classification, based on the epoch with the best validation accuracy) also contributed to better model performances, which matched recent analyses in foundation model training Xue et al. (2024). The dropout rate and weight decay rate also work better with a small value (illustrated in Extended Data Figures 4 (b) and (c)). Meanwhile, we found that increasing the number of sampled cells does not always enhance the performances of PaSCient, shown in Extended Data Figure 4 (d) for the experiments based on binary classification, which also matched previous research about selecting the random sampling policies for multiple instance learning Tarkhan et al. (2023). A suitable range of sampled cell numbers was found to be in $(100, 2000)$.

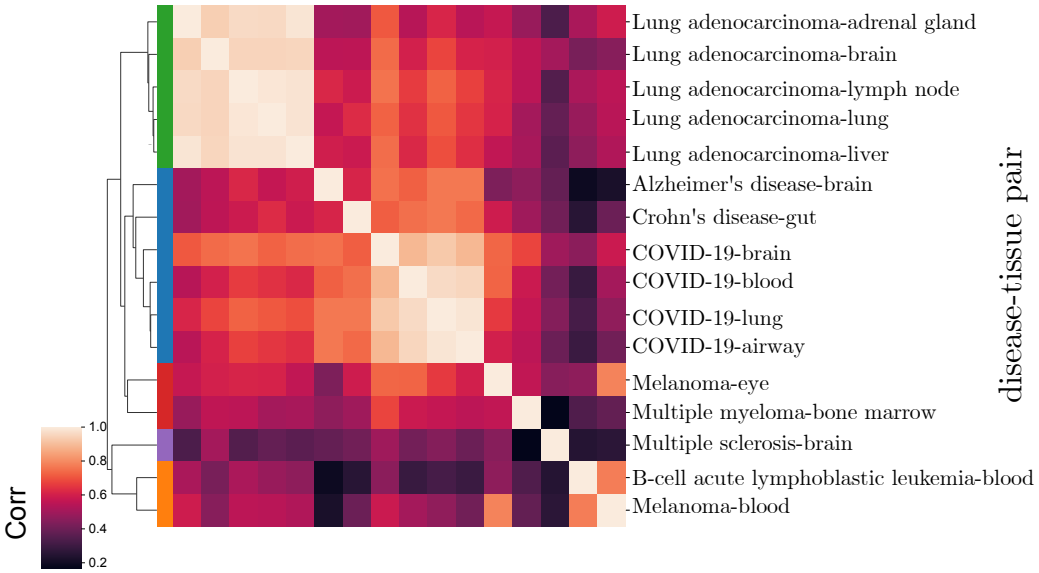
In the multi-class setting, we faced a more complicated condition with different compositions of diseases with different difficulties. Therefore, to investigate the prediction performances specific to different diseases, we hypothesized that model performances was correlated to the frequency of diseases included in the training dataset. PaSCient should be able to classify healthy conditions with diseases whose frequency is low (as multiple binary classification problems). In contrast, PaSCient might face a more and more challenging problem when introducing more diseases in the classification problem (as multiple multi-label classification problems). To validate the first assumption, we



Extended Data Fig. 4: Sensitivity analysis of PaSCient. (**a**) Performance of PaSCient under different learning rates for binary classification. (**b**) Performance of PaSCient under different dropout rates for binary classification. (**c**) Performance of PaSCient under different weight decay rights for binary classification. (**d**) Performance of PaSCient under different number of sampled cells for binary classification. (**e**) Performances of PaSCient under the cumulative setting of different diseases states. (**f**) Performance of PaSCient for classifying healthy samples and other samples with different diseases. (**g**) Performance of PaSCient under different widths of neural network layers. (**h**) Performance of PaSCient under different depths of neural network layers.

separated the eight diseases into eight classification problems versus healthy conditions and trained PaSCient to distinguish them. The results of this experiment are presented in Extended Data Figure 4 (e). We found that diseases with lower frequency were generally easier to identify. Meanwhile, we added diseases to evaluate the cumulative performances from the largest frequency to the lowest frequency shown in Extended Data Figure 4 (f), which showed that introducing more diseases might reduce classification performances.

To explore the scaling law of PaSCient, we considered adjusting two architecture parameters, including (1) the width of layers (shown in Extended Data Figure 4 (g)) and (2) the number of layers (shown in Extended Data Figure 4 (h)). We found that increasing the width of the layers can help improving the prediction performance, while increasing the number of layers might harm the prediction performance.



Extended Data Fig. 5: Demonstration of patient-sample-level embeddings. Correlation coefficients of disease embeddings across different tissues. The clusters are generated based on the hierarchical clustering with one minus correlation coefficient as distance.

H PaSCient learns the differences and similarities of different diseases in the representation space

Modeling patients is a complex problem, so using classifier performance to measure patient representation is insufficient to show that we have learned meaningful patient embeddings. For the given group of patient embeddings, We can categorize them by disease as well as by tissue of their origin to study the effects of the same disease on different tissues. By averaging the patient embeddings by both diseases and tissues and computing the correlation matrix, we visualized the correlation results in Figure 5. We found that disease embeddings from COVID-19 and lung adenocarcinoma have a high correlation across different tissues, which implied that PaSCient successfully learned the patient sample representations across different tissues from the same disease. Our embeddings also captured signals of different tissues within the same disease demonstrated by the results of hierarchical clustering. Such findings could also be supported by recent research about the multi-tissue damages of COVID-19 Chen et al. (2020) and lung adenocarcinoma Zhu et al. (2021), as these diseases tended to affect different tissues jointly.

I Evaluating patient embeddings with large language models

Rigorously evaluating the quality of the embeddings produced by a given method is a challenging task, that requires meta-data annotations that is not always available. To address this challenge, we constructed a disease similarity measure based on the text descriptions of each disease, extracted from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database Kanehisa & Goto (2000); Kanehisa (2019); Kanehisa et al. (2023) and National Center for Biotechnology Information (NCBI) Sayers et al. (2024). Each disease description was converted to an embedding using the OpenAI text-embedding tool OpenAI et al. (2024). Similarity between diseases was then obtained by computing the Pearson Correlation Coefficients (PCCs) between their respective embedding. The resulting similarity matrix is given in Extended Data Figure 6(a).

We then constructed a similarity measure between diseases from the embeddings of PaSCient by averaging the embbedings of all patients with a given disease and computing the pairwise Pearson Correlation Coefficients. Lastly, we used a similar procedure for computing disease similarities from pseudobulk data. The resulting similarity matrices are presented in Extended Data Figure 6(a).

To quantitatively assess the discrepancy between the text-based disease similarities and PaSCient-based similarities, we computed the PCCs between both similarity matrices. We found that the correlation from PaSCient's result (PCC=0.65, p-value=9.5e-33) was higher than the correlation from pseudobulk gene expression levels (PCC=0.28, p-value=2.6e-6), suggesting that PaSCient can learn better patient representations by considering both the differences and similarities across different diseases.

Furthermore, we investigated whether PaSCient could capture the similarity of certain disease-tissue embeddings with other embeddings by visualizing the relationship between the correlation from text embeddings and the correlation from PaSCient (Extended Data Figure 6(b,c)). We found that PaSCient aligned with the human understanding of diseases with patient representations from transcriptomic data for the lung samples with lung cancer (PCC=0.83, p-value=1.2e-4) and for the lung samples with COVID-19 (PCC=0.72, p-value=2.0e-3). Overall, these preliminary results show that text embeddings can be a promising metric for evaluating the reliability of learned patient representations.

J Explainability of multi-disease states

We visualized the importance scores across cell states and genes by different diseases in Extended Data Figure 7 and Extended Data Figure 8.

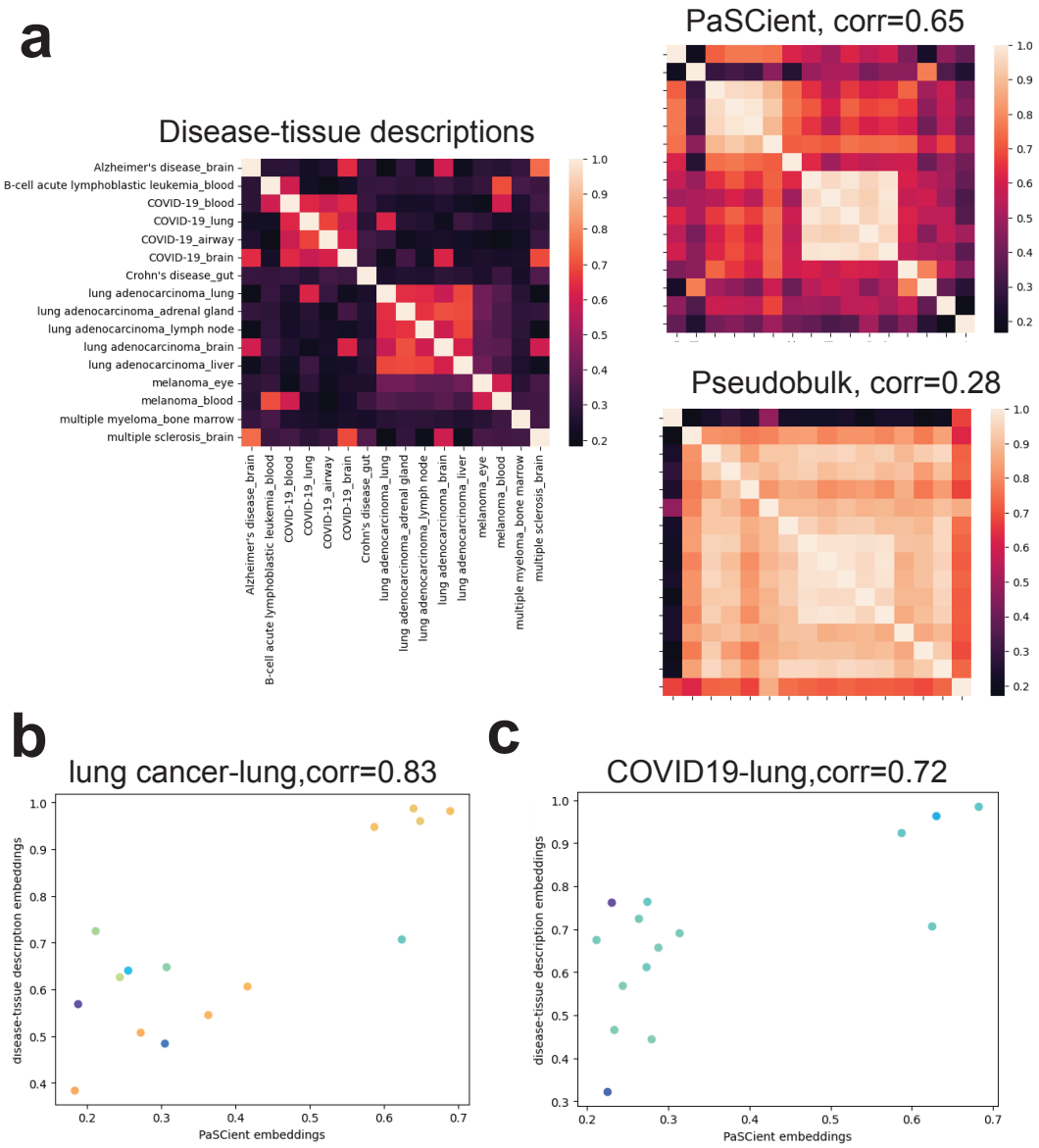
K Association of cell-types importance scores and COVID-19 severity

We visualized the association between attribution scores and COVID-19 severity for each cell-type in Extended Data Figure 9.

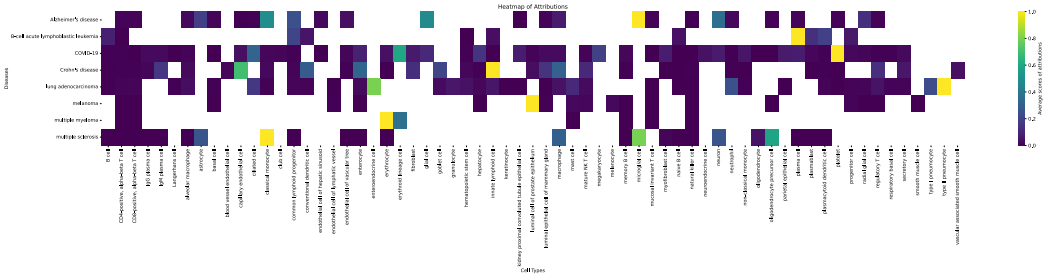
L Case study: Identification of treatment responders in melanoma

To complement our quality-assessment of the patient representations learnt by PaSCient, we investigated whether it could identify treatment responders for T-cell immunotherapy in melanoma. We collected two scRNA-seq datasets from patients with melanoma treated with T-cell immunotherapy, and for which a binary treatment outcome label was available Sade-Feldman et al. (2018); Yost et al. (2019).

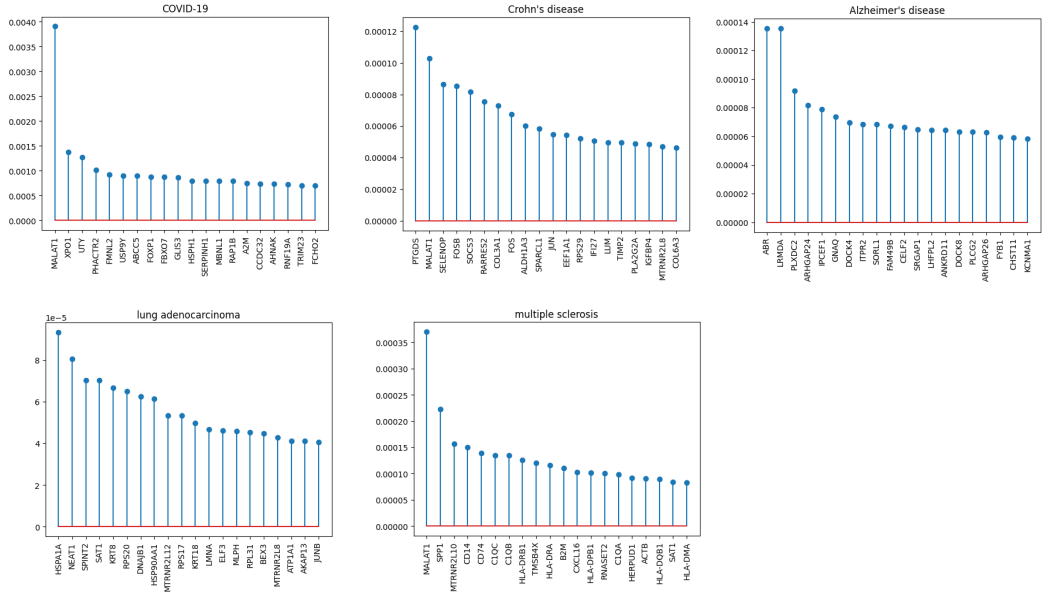
Following the settings of Lin & Sun (2024), we utilized the former dataset as a training dataset and the latter as a testing dataset. For reference, we used the pseudobulk data from original expression profiles by patients with a Support Vector Classifier (SVC), and extracted the patient representations by querying PaSCient with the original gene expression profiles and performed classification under the same training/testing datasets. The whole workflow is shown in Extended Data Figure 10 (a). We also visualized the sample embeddings of training dataset in Extended Data Figure 10 (b) and observed clusters for non-response samples. Furthermore, we visualized the classification results in Extended Data Figure 10 (c), and the SVC using representations form PaSCient as input could outperform the baseline model under different classification metrics, especially because the baseline model predicted all patient samples as drug-responsible.



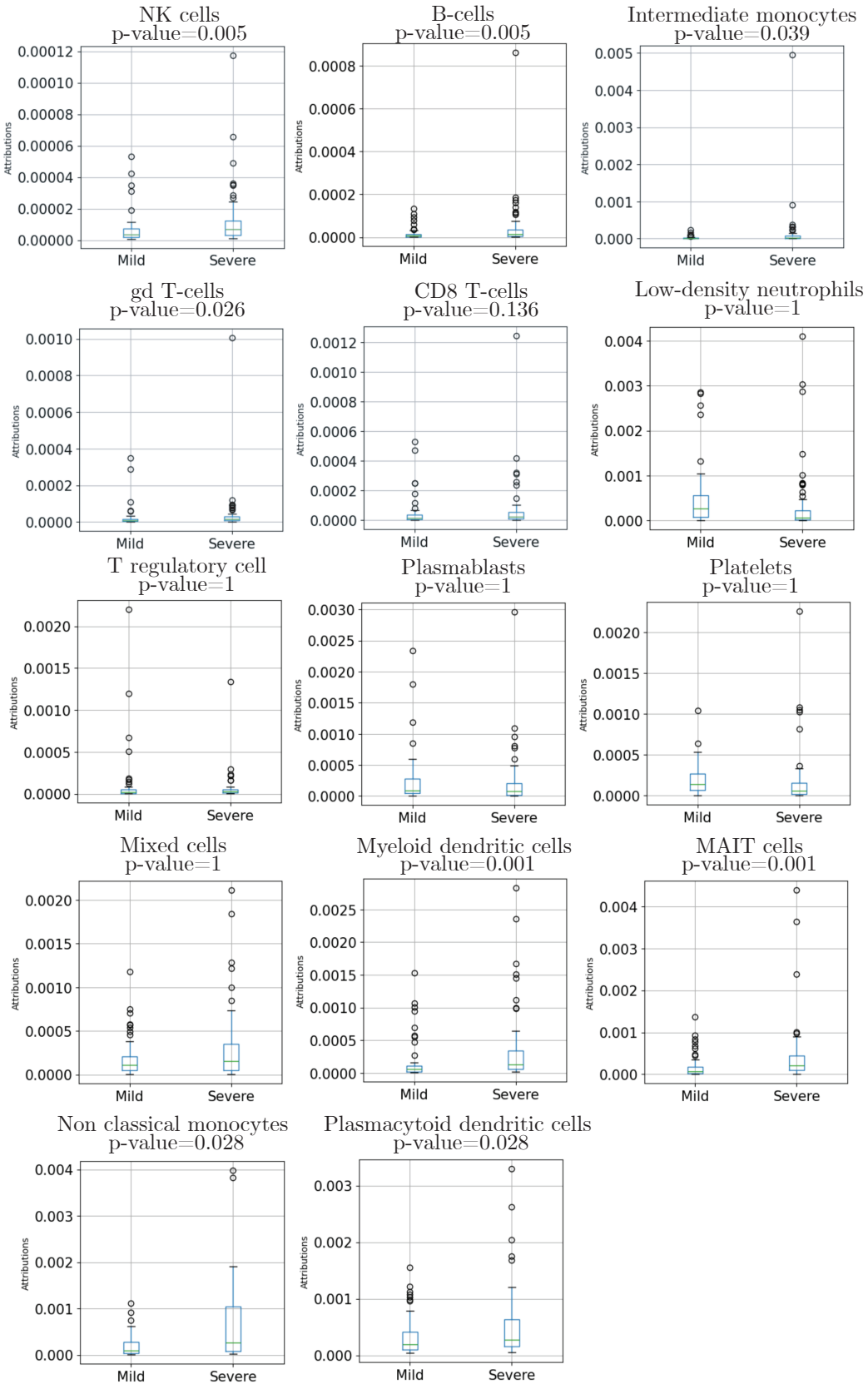
Extended Data Fig. 6: Evaluation patient embeddings with the prior information from text descriptions. (**a**): Correlation matrices of embeddings computed based on different rules for disease similarity across different diseases. The sources and correlations between computed matrix and ground truth matrix are labelled in the figure. (**b**) Scatter plot between the correlation computed with embeddings from PaSCient and the correlation computed with text embeddings for describing the similarity of lung cancer in lung and other disease-tissue pairs. (**c**) Scatter plot between the correlation computed with embeddings from PaSCient and the correlation computed with text embeddings for describing the similarity of COVID19 in lung and other disease-tissue pairs. All the p-values corresponding to correlations are significant.



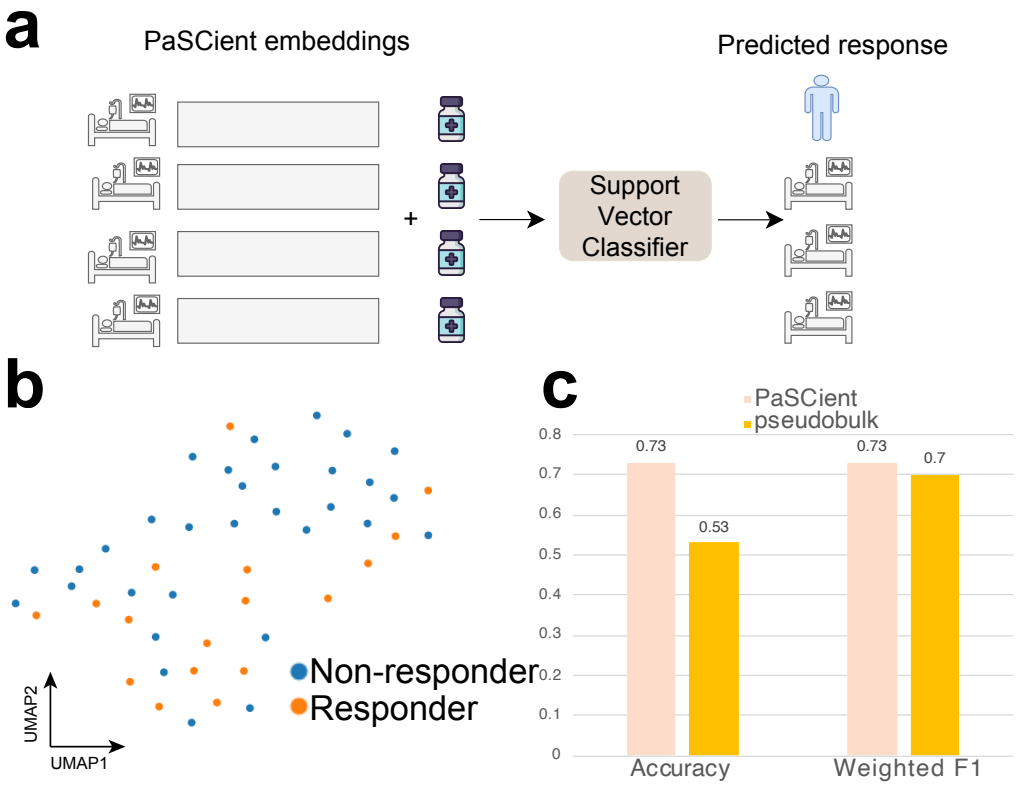
Extended Data Fig. 7: The averaged attributions of different cell types for diseases. The attributions are scaled for each diseases for better visualization, and the cell types are corrected and annotated with SCimilarity.



Extended Data Fig. 8: Averaged attributions of each gene for different diseases. Each figure represents one disease and we select top 20 genes to present, ranked by their average attributions.



Extended Data Fig. 9: Results of t-tests for the association between attribution scores and COVID-19 severity for each cell-type. p-values are Bonferroni corrected.



Extended Data Fig. 10: Performances of PaSCient for patient-level treatment response prediction. (**a**): Overview of treatment prediction task. We utilize patient sample embeddings from PaSCient as input and classify these samples with their known treatment responses and transfer the knowledge to predict unknown treatment responses in the testing dataset. (**b**): Visualization of sample embeddings colored by treatment responses. (**c**): Benchmarking results between PaSCient embeddings and pseudobulk gene expression levels as inputs for this task.