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], and 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
*Equal contribution
✉ represents co-corresponding authors

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.

Keywords: Disease Modelling, Multi-Instance Learning, Single-Cell Transcriptomics, Foundation Model

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 [10,3,41,37]. 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 [45]. The large number of cell profiles within and across experiments has opened the way to discoveries from new cell types [23], distinct genes programs associated with response to therapy or drug resistance, specific marker genes [44,35], and unique patient subsets [46,53]. 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 [29,49,19,20,11,1]. 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 [4,5]. 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 [16,36,59,39], 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 [34]. 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 [31], the imbalanced composition of different tissues, cell types, and diseases [13], and the noise of scRNA-seq data [21,8].

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 [55,6] 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 [52] 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 [63]. 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) [52]. 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 [5]. 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 ∼9% of the samples, while multiple sclerosis only for ∼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 [2,14]. We compare our

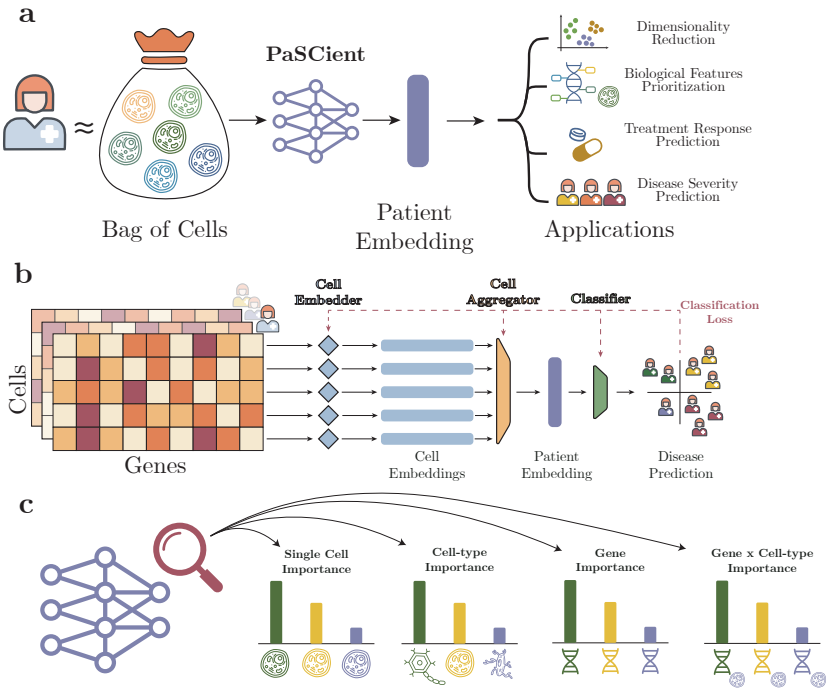

Fig. 1. The landscape of PaSCient. (**a**) Model description and applications. PaSCient abstracts each patient as a bag of cells and outputs a single vector summarizing the patient's cellular context. This vector can be used for various downstream tasks, such as dimensionality reduction, visualization, 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.

approach with different embedding baselines, such as a simple pseudo-bulk approach, using cell-type proportions (CTP), as well as state-of-the-art single-cell foundation models (CellPLM [56] and SCimilarity [17]). For each of these methods, we consider two classifiers to predict the label from the patient embedding: k-Nearest Neighbor Classifier (kNN) [43] 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, in-

cluding the comparison with the most recent domain-expert model ScRAT [36], 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.

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.

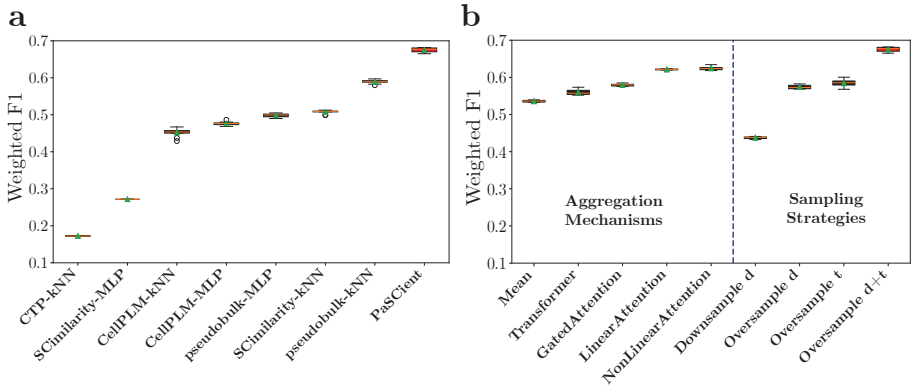

Fig. 2. Benchmarking the performance of PaSCient on multi-disease classification. (**a**) Weighted F1-score results. Performance comparison between PaSCient and relevant baseline models, with standard deviations calculated from experiments using different seeds. PaSCientemploys non-linear attention aggregation 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).

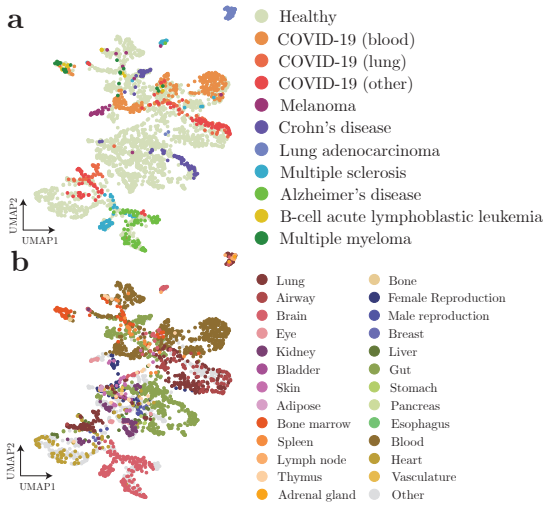

Fig. 3. Patient embeddings using PaSCient organize by both tissue and disease. Uniform manifold approximation and projection (UMAP) of patient embeddings colored by each of 8 most common disease labels (**a**) or by tissue (**b**). We only visualize the samples whose disease-state labels exist in all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

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.

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 [24,58].

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 [38,60,51,64,12].

We validate the set of important genes uncovered by PaSCient by measuring the overlap with the set of differentially expressed genes from ToppCell [22].

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.

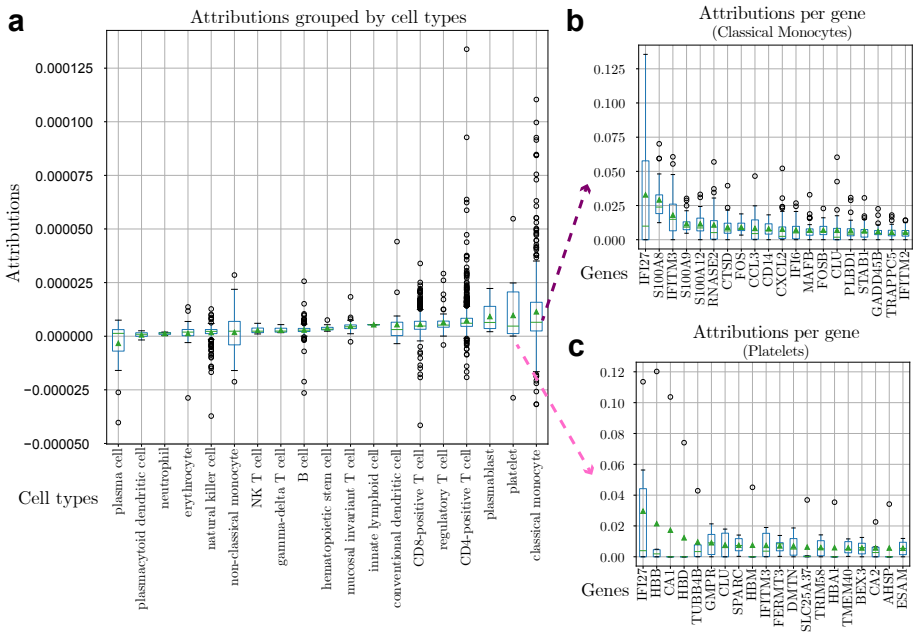

Fig. 4. Prioritizing cell types and cell-type-specific genes for COVID-19 by integrated gradients (IG) analysis. (**a**) Attributions averaged over cells and genes for each cell type (each point is a patient). Cell types are ranked by their mean attribution, with classical monocytes and platelets identified 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.

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) [32,50,57,30], 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.

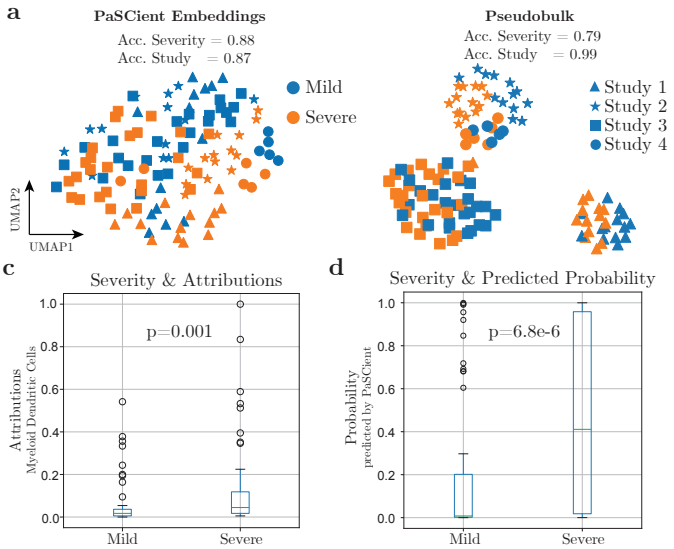

Fig. 5. PaSCient captures disease severity in COVID-19 patients. (**a**) Patient embeddings generated by PaSCient and PCA on pseudo-bulk data, colored by disease severity. PaSCient organizes patient representations based on disease severity, whereas the pseudo-bulk embeddings are influenced by study-specific 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 [9,17,15], and multi-cellular representations models [16,36,59] 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 [52]. 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

1. Responsible use of polygenic risk scores in the clinic: potential benefits, risks and gaps. Nature medicine **27**(11), 1876–1884 (2021)
2. Abdelaal, T., Michielsen, L., Cats, D., Hoogduin, D., Mei, H., Reinders, M.J., Mahfouz, A.: A comparison of automatic cell identification methods for single-cell rna sequencing data. Genome biology **20**, 1–19 (2019)
3. Arowoogun, J.O., Babawarun, O., Chidi, R., Adeniyi, A.O., Okolo, C.A.: 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)
4. Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M., et al.: Ncbi geo: archive for functional genomics data sets—update. Nucleic acids research **41**(D1), D991–D995 (2012)
5. Biology, C.S.C., Abdulla, S., Aevermann, B., Assis, P., Badajoz, S., Bell, S.M., Bezzi, E., Cakir, B., Chaffer, J., Chambers, S., et al.: Cz cellxgene discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. BioRxiv pp. 2023–10 (2023)
6. Boyeau, P., Hong, J., Gayoso, A., Kim, M., McFaline-Figueroa, J.L., Jordan, M.I., Azizi, E., Ergen, C., Yosef, N.: Deep generative modeling of sample-level heterogeneity in single-cell genomics. BioRxiv pp. 2022–10 (2022)
7. Chen, Y.M., Zheng, Y., Yu, Y., Wang, Y., Huang, Q., Qian, F., Sun, L., Song, Z.G., Chen, Z., Feng, J., et al.: Blood molecular markers associated with covid-19 immunopathology and multi-organ damage. The EMBO journal **39**(24), e105896 (2020)
8. Chu, S.K., Zhao, S., Shyr, Y., Liu, Q.: Comprehensive evaluation of noise reduction methods for single-cell rna sequencing data. Briefings in bioinformatics **23**(2), bbab565 (2022)
9. Cui, H., Wang, C., Maan, H., Pang, K., Luo, F., Duan, N., Wang, B.: scgpt: toward building a foundation model for single-cell multi-omics using generative ai. Nature Methods pp. 1–11 (2024)
10. Dash, S., Shakyawar, S.K., Sharma, M., Kaushik, S.: Big data in healthcare: management, analysis and future prospects. Journal of big data **6**(1), 1–25 (2019)
11. Dattani, S., Howard, D.M., Lewis, C.M., Sham, P.C.: Clarifying the causes of consistent and inconsistent findings in genetics. Genetic epidemiology **46**(7), 372–389 (2022)
12. Deniz, S., Uysal, T.K., Capasso, C., Supuran, C.T., Ozensoy Guler, O.: 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)
13. Ferretti, M.T., Iulita, M.F., Cavedo, E., Chiesa, P.A., Schumacher Dimech, A., Santuccione Chadha, A., Baracchi, F., Girouard, H., Misoch, S., Giacobini, E., et al.: Sex differences in alzheimer disease—the gateway to precision medicine. Nature Reviews Neurology **14**(8), 457–469 (2018)
14. Grandini, M., Bagli, E., Visani, G.: Metrics for multi-class classification: an overview. arXiv preprint arXiv:2008.05756 (2020)
15. Hao, M., Gong, J., Zeng, X., Liu, C., Guo, Y., Cheng, X., Wang, T., Ma, J., Zhang, X., Song, L.: Large-scale foundation model on single-cell transcriptomics. Nature Methods pp. 1–11 (2024)

16. He, B., Thomson, M., Subramaniam, M., Perez, R., Ye, C.J., Zou, J.: Cloudpred: Predicting patient phenotypes from single-cell rna-seq pp. 337–348 (2021)
17. Heimberg, G., Kuo, T., DePianto, D., Heigl, T., Diamant, N., Salem, O., Scalia, G., Biancalani, T., Turley, S., Rock, J., et al.: Scalable querying of human cell atlases via a foundational model reveals commonalities across fibrosis-associated macrophages. bioRxiv pp. 2023–07 (2023)
18. Hernandez, D., Kaplan, J., Henighan, T., McCandlish, S.: Scaling laws for transfer. arXiv preprint arXiv:2102.01293 (2021)
19. Hong, H., Xu, L., Su, Z., Liu, J., Ge, W., Shen, J., Fang, H., Perkins, R., Shi, L., Tong, W.: Pitfall of genome-wide association studies: Sources of inconsistency in genotypes and their effects (2012)
20. Ioannidis, J.P.: Non-replication and inconsistency in the genome-wide association setting. Human heredity **64**(4), 203–213 (2007)
21. Janssen, P., Kliesmete, Z., Vieth, B., Adiconis, X., Simmons, S., Marshall, J., McCabe, C., Heyn, H., Levin, J.Z., Enard, W., et al.: The effect of background noise and its removal on the analysis of single-cell expression data. Genome Biology **24**(1), 140 (2023)
22. Jin, K., Bardes, E.E., Mitelpunkt, A., Wang, J.Y., Bhatnagar, S., Sengupta, S., Krummel, D.P., Rothenberg, M.E., Aronow, B.J.: A web portal and workbench for biological dissection of single cell covid-19 host responses. bioRxiv pp. 2021–06 (2021)
23. Jindal, A., Gupta, P., Jayadeva, Sengupta, D.: Discovery of rare cells from voluminous single cell expression data. Nature communications **9**(1), 4719 (2018)
24. Junqueira, C., Crespo, Â., Ranjbar, S., De Lacerda, L.B., Lewandrowski, M., Ingber, J., Parry, B., Ravid, S., Clark, S., Schrimpf, M.R., et al.: Fcγr-mediated sars-cov-2 infection of monocytes activates inflammation. Nature **606**(7914), 576–584 (2022)
25. Kanehisa, M.: Toward understanding the origin and evolution of cellular organisms. Protein Science **28**(11), 1947–1951 (2019)
26. Kanehisa, M., Furumichi, M., Sato, Y., Kawashima, M., Ishiguro-Watanabe, M.: Kegg for taxonomy-based analysis of pathways and genomes. Nucleic acids research **51**(D1), D587–D592 (2023)
27. Kanehisa, M., Goto, S.: Kegg: kyoto encyclopedia of genes and genomes. Nucleic acids research **28**(1), 27–30 (2000)
28. Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D.: Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 (2020)
29. Kau, A.L., Korenblat, P.E.: 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)
30. Lee, J.S., Park, S., Jeong, H.W., Ahn, J.Y., Choi, S.J., Lee, H., Choi, B., Nam, S.K., Sa, M., Kwon, J.S., et al.: Immunophenotyping of covid-19 and influenza highlights the role of type i interferons in development of severe covid-19. Science immunology **5**(49), eabd1554 (2020)
31. Leek, J.T., Scharpf, R.B., Bravo, H.C., Simcha, D., Langmead, B., Johnson, W.E., Geman, D., Baggerly, K., Irizarry, R.A.: Tackling the widespread and critical impact of batch effects in high-throughput data. Nature Reviews Genetics **11**(10), 733–739 (2010)
32. Lemsara, A., Chan, A., Wolff, D., Marschollek, M., Li, Y., Dieterich, C.: Robust machine learning predicts covid-19 disease severity based on single-cell rna-seq from multiple hospitals. medRxiv pp. 2022–10 (2022)

33. Lin, Z., Sun, W.: Supervised deep learning with gene annotation for cell classification. bioRxiv pp. 2024–07 (2024)
34. Litinetskaya, A., Shulman, M., Hediyeh-zadeh, S., Moinfar, A.A., Curion, F., Szalata, A., Omidi, A., Lotfollahi, M., Theis, F.J.: Multimodal weakly supervised learning to identify disease-specific changes in single-cell atlases. bioRxiv pp. 2024–07 (2024)
35. Liu, T., Long, W., Cao, Z., Wang, Y., He, C.H., Zhang, L., Strittmatter, S.M., Zhao, H.: Cosgenegate selects multi-functional and credible biomarkers for single-cell analysis. bioRxiv pp. 2024–05 (2024)
36. Mao, Y., Lin, Y.Y., Wong, N.K., Volik, S., Sar, F., Collins, C., Ester, M.: Phenotype prediction from single-cell rna-seq data using attention-based neural networks. Bioinformatics p. btae067 (2024)
37. Marx, V.: The big challenges of big data. Nature **498**(7453), 255–260 (2013)
38. Mellett, L., Khader, S.A.: S100a8/a9 in covid-19 pathogenesis: Impact on clinical outcomes. Cytokine & Growth Factor Reviews **63**, 90–97 (2022)
39. Mitchel, J., Gordon, M.G., Perez, R.K., Biederstedt, E., Bueno, R., Ye, C.J., Kharchenko, P.V.: Coordinated, multicellular patterns of transcriptional variation that stratify patient cohorts are revealed by tensor decomposition. Nature Biotechnology pp. 1–10 (2024)
40. Musgrave, K., Belongie, S.J., Lim, S.N.: Pytorch metric learning. ArXiv **abs/2008.09164** (2020)
41. Obermeyer, Z., Emanuel, E.J.: Predicting the future—big data, machine learning, and clinical medicine. New England Journal of Medicine **375**(13), 1216–1219 (2016)
42. OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., Bello, I., Berdine, J., Bernadett-Shapiro, G., Berner, C., Bogdonoff, L., Boiko, O., Boyd, M., Brakman, A.L., Brockman, G., Brooks, T., Brundage, M., Button, K., Cai, T., Campbell, R., Cann, A., Carey, B., Carlson, C., Carmichael, R., Chan, B., Chang, C., Chantzis, F., Chen, D., Chen, S., Chen, R., Chen, J., Chen, M., Chess, B., Cho, C., Chu, C., Chung, H.W., Cummings, D., Currier, J., Dai, Y., Decareaux, C., Degry, T., Deutsch, N., Deville, D., Dhar, A., Dohan, D., Dowling, S., Dunning, S., Ecoffet, A., Eleti, A., Eloundou, T., Farhi, D., Fedus, L., Felix, N., Fishman, S.P., Forte, J., Fulford, I., Gao, L., Georges, E., Gibson, C., Goel, V., Gogineni, T., Goh, G., Gontijo-Lopes, R., Gordon, J., Grafstein, M., Gray, S., Greene, R., Gross, J., Gu, S.S., Guo, Y., Hallacy, C., Han, J., Harris, J., He, Y., Heaton, M., Heidecke, J., Hesse, C., Hickey, A., Hickey, W., Hoeschele, P., Houghton, B., Hsu, K., Hu, S., Hu, X., Huizinga, J., Jain, S., Jain, S., Jang, J., Jiang, A., Jiang, R., Jin, H., Jin, D., Jomoto, S., Jonn, B., Jun, H., Kaftan, T., Łukasz Kaiser, Kamali, A., Kanitscheider, I., Keskar, N.S., Khan, T., Kilpatrick, L., Kim, J.W., Kim, C., Kim, Y., Kirchner, J.H., Kiros, J., Knight, M., Kokotajlo, D., Łukasz Kondraciuk, Kondrich, A., Konstantinidis, A., Kosic, K., Krueger, G., Kuo, V., Lampe, M., Lan, I., Lee, T., Leike, J., Leung, J., Levy, D., Li, C.M., Lim, R., Lin, M., Lin, S., Litwin, M., Lopez, T., Lowe, R., Lue, P., Makanju, A., Malfacini, K., Manning, S., Markov, T., Markovski, Y., Martin, B., Mayer, K., Mayne, A., McGrew, B., McKinney, S.M., McLeavey, C., McMillan, P., McNeil, J., Medina, D., Mehta, A., Menick, J., Metz, L., Mishchenko, A., Mishkin, P., Monaco, V., Morikawa, E., Mossing, D., Mu, T., Murati, M., Murk, O., Mély, D., Nair, A., Nakano, R., Nayak, R., Neelakantan, A., Ngo, R., Noh, H., Ouyang, L., O'Keefe, C., Pachocki, J., Paino, A., Palermo, J., Pantuliano, A., Parascandolo, G., Parish, J., Parparita, E.,

Passos, A., Pavlov, M., Peng, A., Perelman, A., de Avila Belbute Peres, F., Petrov, M., de Oliveira Pinto, H.P., Michael, Pokorny, Pokrass, M., Pong, V.H., Powell, T., Power, A., Power, B., Proehl, E., Puri, R., Radford, A., Rae, J., Ramesh, A., Raymond, C., Real, F., Rimbach, K., Ross, C., Rotsted, B., Roussez, H., Ryder, N., Saltarelli, M., Sanders, T., Santurkar, S., Sastry, G., Schmidt, H., Schnurr, D., Schulman, J., Selsam, D., Sheppard, K., Sherbakov, T., Shieh, J., Shoker, S., Shyam, P., Sidor, S., Sigler, E., Simens, M., Sitkin, J., Slama, K., Sohl, I., Sokolowsky, B., Song, Y., Staudacher, N., Such, F.P., Summers, N., Sutskever, I., Tang, J., Tezak, N., Thompson, M.B., Tillet, P., Tootoonchian, A., Tseng, E., Tuggle, P., Turley, N., Tworek, J., Uribe, J.F.C., Vallone, A., Vijayvergiya, A., Voss, C., Wainwright, C., Wang, J.J., Wang, A., Wang, B., Ward, J., Wei, J., Weinmann, C., Welihinda, A., Welinder, P., Weng, J., Weng, L., Wiethoff, M., Willner, D., Winter, C., Wolrich, S., Wong, H., Workman, L., Wu, S., Wu, J., Wu, M., Xiao, K., Xu, T., Yoo, S., Yu, K., Yuan, Q., Zaremba, W., Zellers, R., Zhang, C., Zhang, M., Zhao, S., Zheng, T., Zhuang, J., Zhuk, W., Zoph, B.: Gpt-4 technical report (2024), https://arxiv.org/abs/2303.08774

43. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: Machine learning in python. Journal of machine learning research **12**(Oct), 2825–2830 (2011)
44. Pullin, J.M., McCarthy, D.J.: A comparison of marker gene selection methods for single-cell rna sequencing data. Genome Biology **25**(1), 56 (2024)
45. Regev, A., Teichmann, S.A., Lander, E.S., Amit, I., Benoist, C., Birney, E., Bodenmiller, B., Campbell, P., Carninci, P., Clatworthy, M., et al.: The human cell atlas. elife **6**, e27041 (2017)
46. Rood, J.E., Maartens, A., Hupalowska, A., Teichmann, S.A., Regev, A.: Impact of the human cell atlas on medicine. Nature medicine **28**(12), 2486–2496 (2022)
47. Sade-Feldman, M., Yizhak, K., Bjorgaard, S.L., Ray, J.P., de Boer, C.G., Jenkins, R.W., Lieb, D.J., Chen, J.H., Frederick, D.T., Barzily-Rokni, M., et al.: Defining t cell states associated with response to checkpoint immunotherapy in melanoma. Cell **175**(4), 998–1013 (2018)
48. Sayers, E.W., Beck, J., Bolton, E.E., Brister, J.R., Chan, J., Comeau, D.C., Connor, R., DiCuccio, M., Farrell, C.M., Feldgarden, M., et al.: Database resources of the national center for biotechnology information. Nucleic Acids Research **52**(D1), D33 (2024)
49. Schaid, D.J., Chen, W., Larson, N.B.: From genome-wide associations to candidate causal variants by statistical fine-mapping. Nature Reviews Genetics **19**(8), 491–504 (2018)
50. Schulte-Schrepping, J., Reusch, N., Paclik, D., Baßler, K., Schlickeiser, S., Zhang, B., Krämer, B., Krammer, T., Brumhard, S., Bonaguro, L., et al.: Severe covid-19 is marked by a dysregulated myeloid cell compartment. Cell **182**(6), 1419–1440 (2020)
51. Shojaei, M., Shamshirian, A., Monkman, J., Grice, L., Tran, M., Tan, C.W., Teo, S.M., Rodrigues Rossi, G., McCulloch, T.R., Nalos, M., et al.: Ifi27 transcription is an early predictor for covid-19 outcomes, a multi-cohort observational study. Frontiers in Immunology **13**, 1060438 (2023)
52. Sundararajan, M., Taly, A., Yan, Q.: Axiomatic attribution for deep networks (2017)
53. Suvà, M.L., Rheinbay, E., Gillespie, S.M., Patel, A.P., Wakimoto, H., Rabkin, S.D., Riggi, N., Chi, A.S., Cahill, D.P., Nahed, B.V., et al.: Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells. Cell **157**(3), 580–594 (2014)

54. Tarkhan, A., Nguyen, T.K., Simon, N., Dai, J.: Survival prediction via deep attention-based multiple-instance learning networks with instance sampling (2023)
55. Wang, H., Leskovec, J., Regev, A.: Metric mirages in cell embeddings. bioRxiv pp. 2024–04 (2024)
56. Wen, H., Tang, W., Dai, X., Ding, J., Jin, W., Xie, Y., Tang, J.: CellPLM: Pre-training of cell language model beyond single cells (2024), https://openreview.net/forum?id=BKXvPDekud
57. Wilk, A.J., Lee, M.J., Wei, B., Parks, B., Pi, R., Martínez-Colón, G.J., Ranganath, T., Zhao, N.Q., Taylor, S., Becker, W., et al.: Multi-omic profiling reveals widespread dysregulation of innate immunity and hematopoiesis in covid-19. Journal of Experimental Medicine **218**(8), e20210582 (2021)
58. Wool, G.D., Miller, J.L.: The impact of covid-19 disease on platelets and coagulation. Pathobiology **88**(1), 15–27 (2021)
59. Xiong, G., Bekiranov, S., Zhang, A.: Protocell4p: an explainable prototype-based neural network for patient classification using single-cell rna-seq. Bioinformatics **39**(8), btad493 (2023)
60. Xu, F., Wang, G., Zhao, F., Huang, Y., Fan, Z., Mei, S., Xie, Y., Wei, L., Hu, Y., Wang, C., et al.: Ifitm3 inhibits sars-cov-2 infection and is associated with covid-19 susceptibility. Viruses **14**(11), 2553 (2022)
61. Xue, F., Fu, Y., Zhou, W., Zheng, Z., You, Y.: To repeat or not to repeat: Insights from scaling llm under token-crisis. Advances in Neural Information Processing Systems **36** (2024)
62. Yost, K.E., Satpathy, A.T., Wells, D.K., Qi, Y., Wang, C., Kageyama, R., McNamara, K.L., Granja, J.M., Sarin, K.Y., Brown, R.A., et al.: Clonal replacement of tumor-specific t cells following pd-1 blockade. Nature medicine **25**(8), 1251–1259 (2019)
63. Zaheer, M., Kottur, S., Ravanbakhsh, S., Poczos, B., Salakhutdinov, R.R., Smola, A.J.: Deep sets. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/f22e4747da1aa27e363d86d40ff442fe-Paper.pdf
64. Zhang, Z., Lin, F., Liu, F., Li, Q., Li, Y., Zhu, Z., Guo, H., Liu, L., Liu, X., Liu, W., 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)
65. Zhu, S., Ge, T., Hu, J., Jiang, G., Zhang, P.: Prognostic value of surgical intervention in advanced lung adenocarcinoma: a population-based study. Journal of Thoracic Disease **13**(10), 5942 (2021)