Weakly supervised learning uncovers phenotypic signatures in single-cell data

Anastasia Litinetskaya[1,4], Soroor Hediyeh-zadeh[1,2], Amir Ali Moinfar[1,4], Mohammad Lotfollahi[1,3,5], Fabian J. Theis[1,2,3,4†]

1 Institute of Computational Biology, Helmholtz Center Munich, Munich, Germany.

2 School of Life Sciences Weihenstephan, Technical University of Munich, Munich, Germany.

3 Wellcome Sanger Institute, Wellcome Genome Campus, Cambridge, UK.

4 School of Computing, Information and Technology, Technical University of Munich, Munich, Germany.

5 Wellcome–MRC Cambridge Stem Cell Institute, University of Cambridge, Cambridge, UK.

†Correspondence to:

fabian.theis@helmholtz-munich.de

Abstract

To deliver clinically relevant insights from large patient cohorts profiled with single-cell technologies, a key challenge is to relate sample-level and single-cell measurements. We present MultiMIL, a deep learning framework that applies attention-based multiple-instance learning for phenotype prediction and cell state identification. We applied MultiMIL to peripheral blood mononuclear cells from COVID-19 patients, the Human Lung Cell Atlas, and a spatial proteomics breast cancer dataset, demonstrating how our model can be utilized to find phenotype-associated cell states, learn phenotype-informed sample representations, and expand disease signatures.

Keywords

Multiple-instance learning · Phenotype prediction · Cell state identification · Sample representation learning

1 Background

The advancements in single-cell technologies have enabled a consistent increase in the number of cells captured for analysis, as well as the number of samples, i.e., individual donors, tissue biopsies, or other patient-level observations, resulting in larger cohort sizes in atlases [1, 2]. To understand the underlying cellular processes and mechanisms that drive disease phenotypes, a significant challenge remains in linking cell-level signals to patient-level phenotypes in an interpretable manner.

Multiple strategies have been developed for this task, including pseudobulking, which aggregates cell-level data to mimic bulk profiles and facilitates sample-level analysis; methods that identify cell types exhibiting differential transcriptomic signals (e.g., via compositional shifts or marker expression); and direct phenotype prediction models operating at either the cellular or patient level [3–14]. However, these approaches often suffer from limitations. Pseudobulking sacrifices single-cell resolution; many methods are restricted to scRNA-seq and do not support other modalities. Additionally, some methods struggle to disentangle genuine biological variation from technical or batch effects. Even when predictions are made at the sample level, it is not straightforward to link these predictions to the cellular processes driving the disease phenotype.

In machine learning terms, predicting sample-level phenotypes from cell-level data is a weakly supervised learning problem—only sample labels are available, while labels for individual cells are missing. Multiple-instance learning (MIL) offers a suitable framework: each sample (bag) contains multiple cells (instances), and the goal is to learn bag-level labels while identifying which instances are most informative [15]. Attention-based MIL, in particular, enables interpretable aggregation by assigning weights to each cell's contribution to the sample-level decision. MIL has been successfully applied in various biomedical domains, including blood disorder classification from microscopy, learning genetic variant associations with transcriptional states, and T-cell receptor sequence-based cancer detection and phenotype prediction from scRNA-seq data through hierarchical MIL frameworks that incorporate cell-type information [16–19].

When applying MIL to single-cell data, several domain-specific challenges must be considered. Biological signal is often confounded by technical batch effects and measurement noise, complicating the task of distinguishing true disease-associated variation from artifacts. From a machine learning perspective, it is challenging to balance predictive performance with interpretability, as highly predictive models often do not perform robustly enough to rely on for hypothesis generation. For the approach to be helpful in diagnostics or therapeutics, the identification of disease-associated cellular

states must be reproducible, indicating that they drive disease phenotypes rather than reflect noise or artifacts.

To address these limitations, we introduce MultiMIL, a deep MIL-based model for phenotypic prediction and cell prioritization. Since not all cells from diseased samples exhibit disease-induced changes, the MIL approach enables the identification of cells most affected by the disease. Identification is achieved through attention scores, where highly attended cells are most relevant and the least attended cells are the least relevant to the observed phenotype of the sample. This reflects the two main mechanisms — transcriptional shifts or changes in cell type composition — through which the observed phenotypes can be explained. MultiMIL is agnostic to the input modality and can be trained on raw features or latent representations from atlases or foundation models.

We demonstrate the power of MultiMIL's phenotypic prediction for unseen patients and prioritization of disease-specific cell states by analyzing several high-relevance clinical datasets: a human peripheral blood mononuclear cells from COVID-19 patients profiled with single-cell RNA-seq [20], the Human Lung Cell Atlas [21] and a breast cancer dataset, profiled with imaging mass cytometry [22]. We further demonstrate how the disease states identified with MultiMIL help discover novel genes associated with the disease and finally, how the sample representations learned with our model are reflective of the phenotypes. MultiMIL's implementation and reproducibility scripts are available at `http://github.com/theislab/multimil`. We also publish a pipeline for comparing different methods for sample-level prediction from single-cell data at `https://github.com/theislab/sample-prediction-pipeline`, which enables fast, multi-scale analysis of custom datasets.

2 Results

Prioritizing phenotype-specific cells and predicting phenotypes with MultiMIL

MultiMIL is a deep-learning-based model that allows the prediction of sample-level phenotypes from single-cell measurements and the identification of cell states associated with these phenotypes. MultiMIL's model consists of a pooling layer that learns sample-level representations from single-cell representations and a classification network that learns to predict sample-level phenotypes (**Fig. 1**b). We draw inspiration from the multiple-instance learning (MIL) approach [15, 16], where we model samples as bags and cells as instances belonging to a bag. The classification labels are only known on the bag level but not on the instance level, and we are interested in identifying instances associated

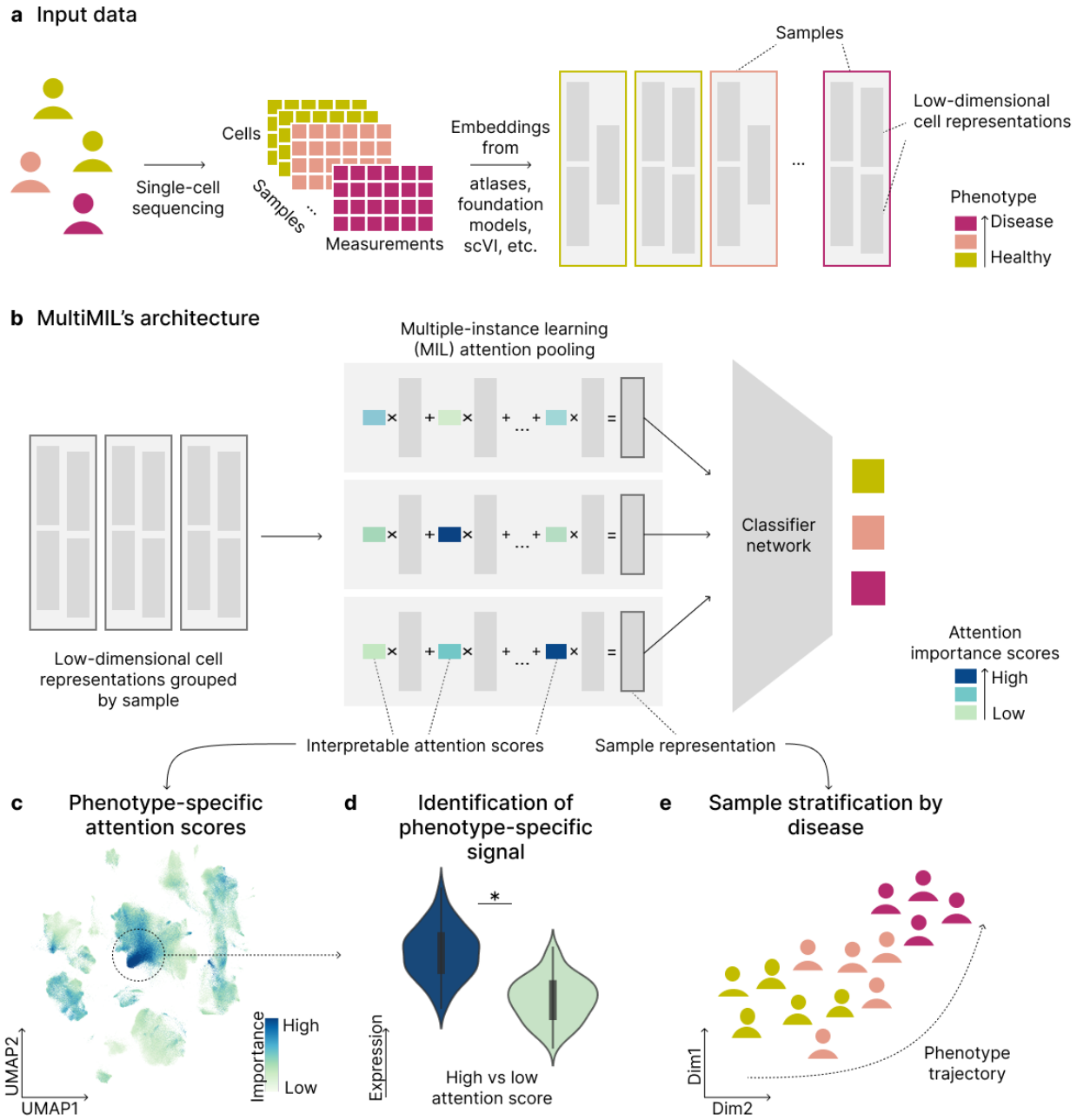


Figure 1 | MultiMIL predicts sample-level phenotypes through interpretable multi-scale learning (a) MultiMIL takes low-dimensional cell representations grouped by sample as input. **(b)** MultiMIL's architecture consists of a MIL attention pooling layer, which takes as input cell representations and outputs sample representations, and a sample classification network that predicts phenotypes from sample representations. **(c)** MultiMIL learns an attention score per cell to identify cell states associated with phenotype and **(d)** allows for novel differential signal identification. **(e)** The learned sample representations reflect the phenotype trajectory.

with the bag label (**Suppl. Fig. 1**).

The input to the model is usually batch-corrected cell embeddings, for instance, embeddings from atlases or foundation models (**Fig. 1**a). It is also possible to train MultiMIL directly on the raw features. The MIL attention pooling layer aggregates the cell-level embeddings into a bag embedding employing attention pooling. During training, the model learns attention weight α_i for each cell i in a bag and then aggregates cell embeddings z_i into a bag representation z_{bag} as a weighted sum $z_{\text{bag}} = \sum_i \alpha_i z_i$. The pooled representation z_{bag} is then fed into a feed-forward classification network that predicts the phenotypes.

MultiMIL provides several ways to interpret the learned attention weights (**Fig. 1**c). Firstly, the higher the weight of a particular cell is, the more important the cell is for the prediction. By learning a score for each cell, we identify and analyze cell states associated with a particular phenotype by selecting cells with high attention scores. This allows for the identification of differential signals between groups at higher resolutions than cell types, as we are able to subset the data to only disease-affected cells (**Fig. 1**d). Attention scores also allow us to learn phenotype-informed sample representations. We can obtain them by averaging the cells' representations with the highest attention scores within a sample (**Fig. 1**e).

MultiMIL is fast to train due to the mini-batch training and the deep-learning nature of the model: the model training for the dataset of approximately 250,000 cells takes less than 6 minutes (Table 3). Computational efficiency, combined with interpretability through attention scores, makes MultiMIL a scalable and practical tool for identifying disease-relevant cellular signals in large single-cell datasets.

Robust phenotype prediction from diverse single-cell embeddings with MultiMIL

In this section, we will investigate how the quality of the embedding affects the prediction performance and the consistency of the learned attention scores. We will also benchmark the model's performance depending on various training and model parameters such as batch size, learning rate or pooling function (the range of tested parameters is shown in Table 1 in Methods). We used a peripheral blood mononuclear cell (PBMC) dataset [20] throughout these experiments. This large-scale dataset consists of 130 healthy and diseased samples and provides annotation on the progression of COVID-19 stages.

First, we investigated how the quality of the embedding and different model parameters affect the performance of MultiMIL. To this end, we obtained low-dimensional representations of the RNA-

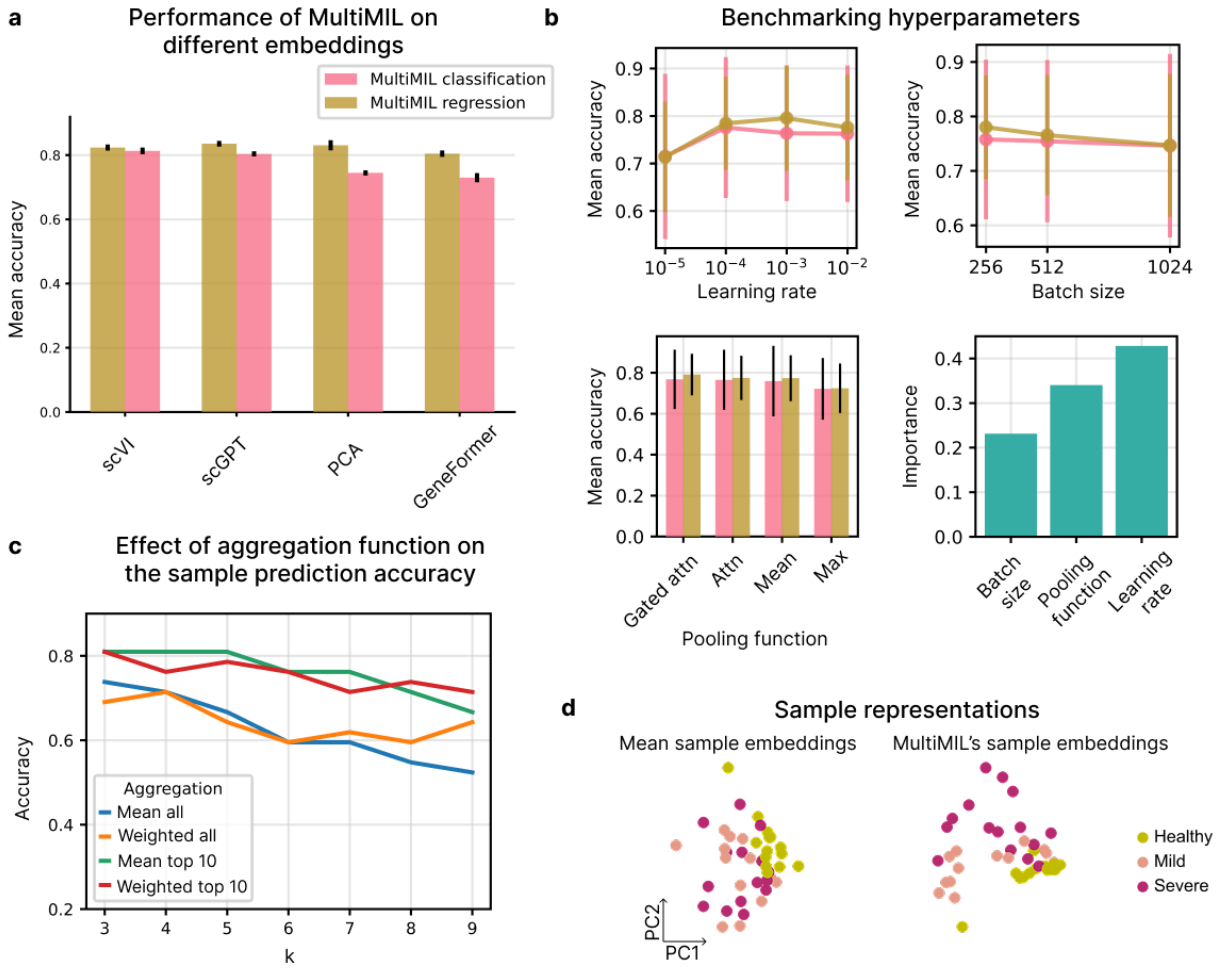


Figure 2 | Benchmarking MultiMIL **(a)** Best accuracy achieved by MultiMIL's classification and regression models on different embeddings. **(b)** Effect of the learning rate, batch size and the pooling function on the accuracy and the feature importance score of these parameters. **(c)** Sample-level kNN prediction accuracy of different sample-representation aggregation methods. **(d)** PC plot of mean sample embeddings and MultiMIL's sample embeddings colored by disease stage.

seq single-cell measurements using scGPT and GeneFormer foundation models. We also trained an scVI model and calculated the PC embeddings before batch correction as two additional baseline embeddings. We subset the data to healthy, mild and severe COVID-19 samples, split the data into five cross-validation splits and train MultiMIL to predict these 3 classes either as a classification task or an ordinal regression task. We observed that the classification accuracy was correlated with the quality of the embedding, which we assessed with scIB integration metrics [23]. The accuracy was the highest for scVI and scGPT embeddings (83% and 84%, respectively, averaged for classification and regression) and lower for PC and GeneFormer embeddings (80% and 79%, respectively) (**Fig. 2**a, **Suppl. Fig. 2**b, c). The regression model consistently outperformed the classification model on the 3-class prediction problem (**Fig. 2**a), and we observed that the variance of the accuracy was also consistently lower for the regression model (**Suppl. Fig. 2**a). However, the regression model performed worse than the classification model on the 6-class task. On the task of binary prediction, classification and regression models performed similarly, indicating that for easier tasks, both models can perform equally well (**Suppl. Fig. 3**d). We additionally compared the default classification and regression models to the alternative architecture where we concatenated one-hot encoded embeddings for sample covariates to the sample embedding before passing it through the classification network (see Methods). Incorporating the covariates decreased the performance accuracy (**Suppl. Fig. 3**e), so we used the model architectures without including the additional sample-level information.

Next, we investigated the effect of the learning rate, batch size and the pooling function on the performance, focusing on the better-performing embeddings, i.e., scVI and scGPT. We also trained a random forest regressor to find which hyperparameters influence the performance most. Gated attention pooling slightly overperformed the mean pooling and the attention pooling, while max pooling consistently underperformed (**Fig. 2**b). Since we want to benefit from the interpretability of the attention scores, we use gated attention as the default for the rest of the analyses. We observed that the batch size did not have a significant effect on the performance, although the overall stability of the predictions was slightly higher for smaller batch sizes (**Suppl. Fig. 2**d). Learning rate was the most important hyperparameter for the accuracy (**Fig. 2**b), so we recommend that users start with a learning rate optimization if the default parameters yield unsatisfactory performance. We also observed that training the model for 50 epochs yielded the best performance, compared to longer training regimes (**Suppl. Fig. 2**e).

We also investigated how consistent the cell attention scores are across cross-validation splits, comparing the results from best-performing models trained on scVI and scGPT embeddings. We ob-

served that the model trained on the scVI embeddings consistently learned high attention for the same cell types (**Suppl. Fig. 3**a,b), so we hypothesize that the dimensionality of the latent space could play a role, where the input of lower dimensionality yields more consistent cell attention scores (30 dimensions for scVI vs. 512 for scGPT).

Finally, we show that by aggregating cells with the highest attention scores, we obtain sample representations most indicative of the disease stages, compared to averaging all cell embeddings or taking a weighted (by attention score) average of the cell embeddings (**Fig. 2**c,d). Averaging cell embedding of the top 10% cells with the highest attention score yielded the best performance, compared to other thresholds (**Suppl. Fig. 3**c).

We showed that MultiMIL's performance benefits from good-quality cell embeddings. The model can robustly predict sample-level covariates across different hyperparameters and aggregation strategies.

MultiMIL accurately predicts COVID-19 stages and identifies disease-associated cell states

Next, we discuss the results on the PBMC dataset [20], comparing MultiMIL's predictive performance to baselines and linking the high-attention cell states to known COVID-19 pathology. We used the scVI embedding subset to healthy, mild and severe COVID-19 samples and selected the best-performing MultiMIL's model on this 3-class prediction task. We assessed the predictive performance of MultiMIL compared to several baseline models and interpreted the cell attention scores (**Fig. 3a**).

In this work, we focus on weakly-supervised prediction where the sample-level labels are known, so we compared our model to supervised baseline models, including a random forest, feed-forward neural network and multiclass regression. Approaches utilizing single-cell data for phenotypic prediction often rely on (pseudo-)bulk data [7, 24], so we included a range of pseudo-bulk baselines in our comparison. Since MIL models generally fall between models that make predictions on the instance (i.e., single-cell) level and models that make predictions on the sample (i.e., bulk) level, we also include cell-level baselines (**Fig. 3b**, Methods). The mean embedding of a sample is the mean of cell embeddings belonging to this sample, and cell-type mean embeddings are calculated as the mean of cell embeddings per cell type and concatenated per sample. Frequency vectors are calculated as relative frequencies of cell types present in each sample. For cell embeddings, the input to the models was the cell embeddings from the scVI space, and the prediction was made for each cell. We note that cell-type mean embeddings and frequency vectors are supervised since the cell type labels are

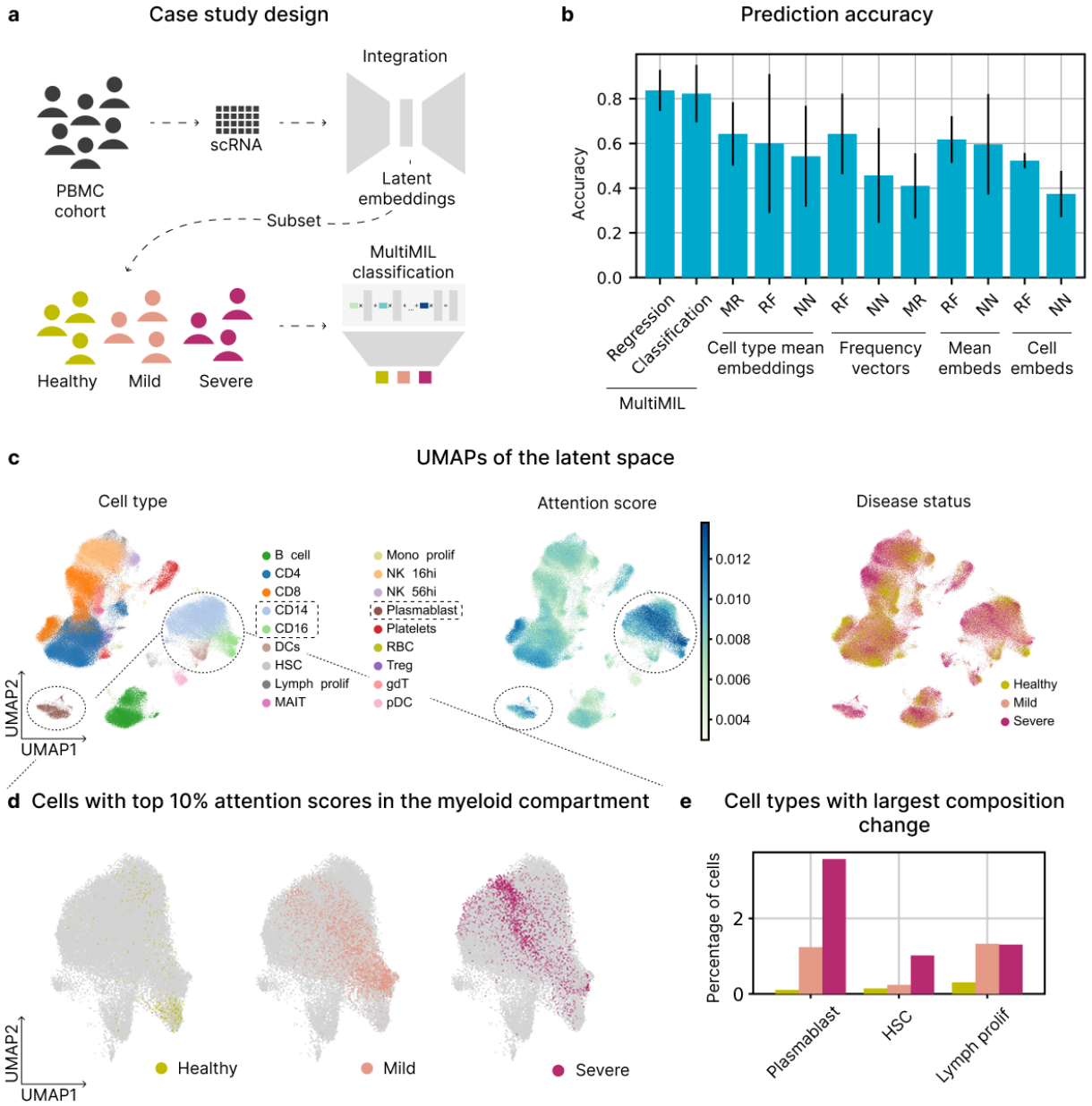


Figure 3 | MultiMIL accurately predicts disease stages on a PBMC dataset. **(a)** Case study design. PBMCs were sequenced with CITE-seq (paired RNA and ADT), integrated to obtain a shared latent representation, subset to healthy, mild and severe COVID-19 samples, and used as input to train MultiMIL's classification network. **(b)** A bar plot showing average accuracies and standard deviations (i.e., the length of an error bar equals two standard deviations) of the five cross-validation runs on the disease-prediction task. MultiMIL was trained in the classification and regression settings. Cell-type mean embeddings and frequency vectors were input to the random forest (RF), feed-forward neural network (NN) and multiclass logistic regression (MR) models. Mean embeddings and cell embeddings were input to the RF and NN models. **(c)** UMAPs of the integrated latent space colored by cell type (left), cell attention scores (middle) and condition (right). The myeloid compartment (i.e., CD14, CD16 monocytes and dendritic cells) and plasmablasts have high attention scores. **(d)** UMAPs of the myeloid compartment showing the healthy, mild and severe COVID-19 cells with the top 10% of attention scores for each condition. **(e)** A bar plot showing the top three cell types with the biggest compositional change from healthy to severe COVID-19, including plasmablasts.

required, while MultiMIL, mean embeddings and cell embeddings are not. MultiMIL outperformed all the baselines in a 5-fold cross-validation experiment, achieving an accuracy of 84% (**Fig. 3b**).

When analyzing diseased samples, we are interested in identifying cell states affected by the disease. By utilizing the attention pooling layer, our model learns a weight for each cell, where higher weights directly correspond to cell states associated with the condition. We only take into account cells with the 10% highest scores per condition, as these cells are most strongly associated with the disease. We observe in **Fig. 3**c that cell types with the highest attention scores are CD14 and CD16 monocytes. Hence, we first examine the myeloid compartment (**Fig. 3**d) and notice a trajectory of highlighted CD14 monocytes from healthy and mild to severe, indicating a shift in expression levels between different stages. Similarly, we find distinct populations of highlighted healthy and mild CD16 monocytes, confirming that the signal learned with MultiMIL aligns with previous studies reporting strong changes in monocytes with the progression of COVID-19 [25, 26].

We also noticed that the plasmablast cluster had a high attention score, so we hypothesized that it could be related to compositional differences. Hence, we next investigated which cell types had the biggest compositional changes between conditions and found that plasmablasts was the cell type most affected between healthy and severe conditions (**Fig. 3**e). MultiMIL identified these compositional changes as indicative of disease progression, also reported in [20]. We additionally ran Milo [13] on the same embeddings and found that cell populations identified by MultiMIL, e.g., CD16 monocytes and platelets, were among the cell types with the highest log-fold-change in composition identified by Milo (**Suppl. Fig. 4**). We note that Milo allows comparisons between two conditions, while MultiMIL identifies condition-specific cell states for multiple classes simultaneously.

These findings reinforce the role of classical and non-classical monocytes in COVID-19 severity progression and demonstrate that MultiMIL can rediscover key immune modulators of disease progression. The ability to distinguish myeloid shifts between mild and severe cases without using cell type labels highlights MultiMIL's utility for studying immune responses.

MultiMIL identifies a subpopulation of IPF-associated macrophages in human lung

Single-cell atlases provide integrated and cell-type-harmonized representations of different systems or organs of interest. These atlases can comprise hundreds of donors, which in turn is crucial to understanding the disease variability and potential therapeutic targets [27]. We demonstrate how MultiMIL can be utilized with existing single-cell atlases. The Human lung cell atlas (HLCA) [21] consists of healthy and diseased donors integrated into a common latent space. We investigated

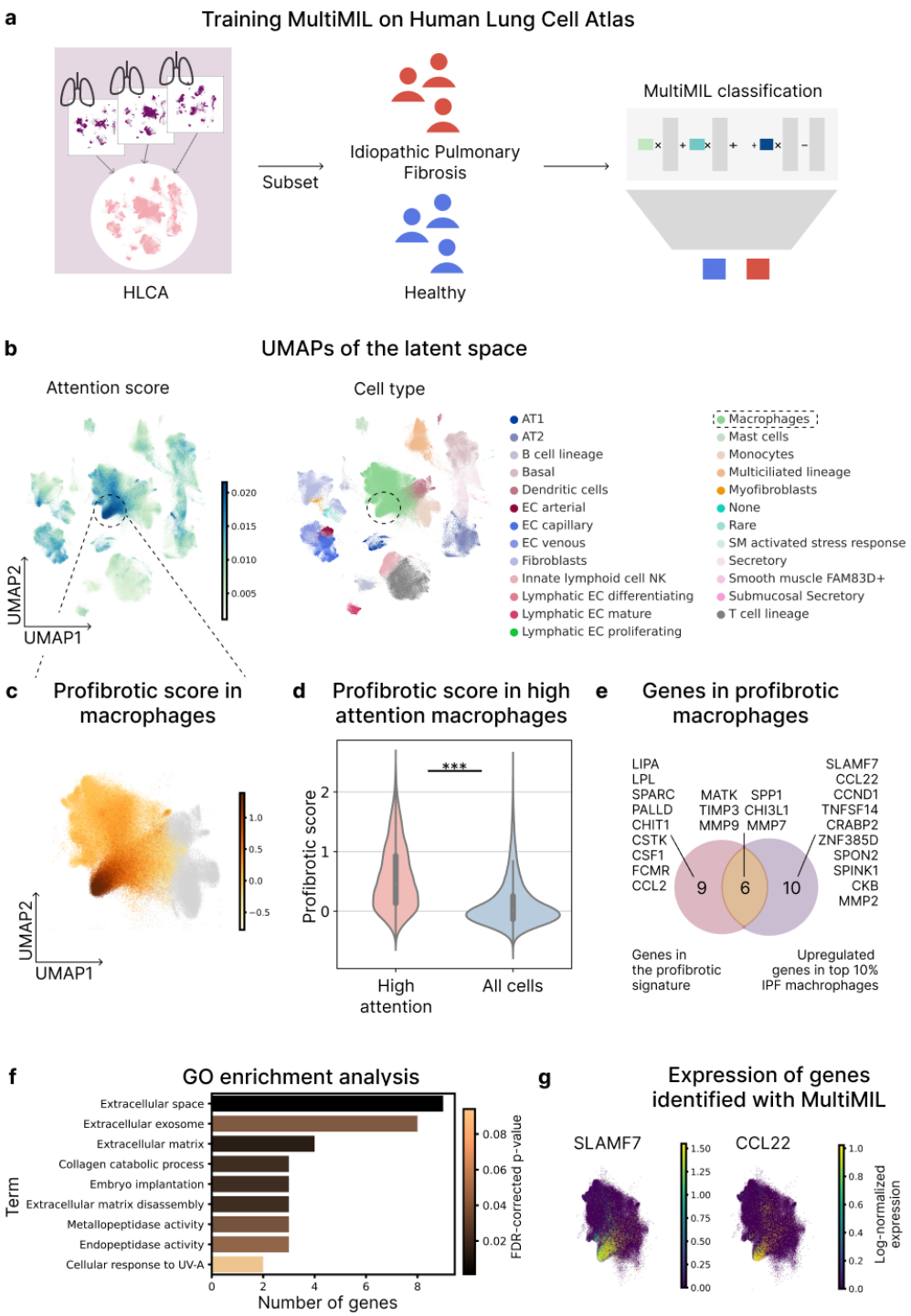


Figure 4 | MultiMIL identifies IPF-associated cell states in human lung macrophages (a) Pre-trained embeddings from the HLCA were subset to healthy and IPF samples and used to train MultiMIL's classification network on the binary classification task. **(b)** UMAPs of the original latent space from the HLCA colored by cell type (top), condition (bottom left), the profibrotic score calculated in macrophages (bottom middle) and cell attention score (bottom right). A subpopulation of macrophages has a high attention score, so we investigate these cells further. **(c)** A UMAP plot of macrophages colored by the profibrotic score. **(d)** Violin plots showing the profibrotic score in high-attention macrophages and all macrophages from IPF patients (p-value<0.001, two-sided t-test). **(e)** A Venn diagram with the genes in the profibrotic signature, the number of genes that are upregulated in the high-attention macrophages compared with all macrophages from IPF patients and the number of genes in the intersection of the two sets. **(f)** GO enrichment analysis of the upregulated genes in the high-attention macrophages. **(g)** UMAPs of the macrophages colored by the expression of *SLAMF7 and CCL22*.

idiopathic pulmonary fibrosis (IPF) and compared diseased and healthy samples. We selected healthy and IPF donors from the atlas and trained MultiMIL in a 5-fold cross-validation setting (**Fig. 4a**). On this classification task, MultiMIL outperformed all baselines, although several other models also achieved strong performance, with accuracies exceeding 90% (**Suppl. Fig. 5d**). Notably, if the primary goal is binary classification rather than interpretability, simple mean-embedding approaches can still offer competitive results.

We examine the learned cell attention scores to analyze which cell states the model learns to associate with the disease. We notice that a subset of macrophages has the highest scores (**Fig. 4b**), so we first show that MultiMIL identifies a subpopulation of $SPP1^{\text{hi}}$ IPF-specific macrophages [5, 28] (**Suppl. Fig. 5a**). We hypothesize that this subpopulation corresponds to profibrotic macrophage populations reported in previous studies [29, 30]. To confirm, we calculate the profibrotic score based on the profibrotic signature introduced in [29] (**Fig. 4c**). We select macrophages from IPF patients and show that the cells with the highest attention score (top 10%) have a significantly higher profibrotic score than all IPF macrophages (**Fig. 4d**). MultiMIL also identifies a KRT17^{+} subpopulation of basal cells (**Suppl. Fig. 5b**) that previously have been reported to be associated with IPF [5, 31].

Cells with high attention can also be used for novel gene signature discovery or to expand the existing signatures. We demonstrate how to identify the gene signature of the IPF-associated macrophage subpopulation using only the attention scores and not relying on previous knowledge. We ran edgeR [12] to find differentially expressed genes between IPF macrophages with the top 10% highest attention scores and all IPF macrophages and identified 16 significantly upregulated genes. Comparing these 16 genes with the genes from the profibrotic signature, we find the overlap of 9 (out of 15) genes (**Fig. 4e**).

The genes identified solely from MultiMIL's high attention group include *SLAMF7*, which has been previously reported to regulate the immune response in lung macrophages during polymicrobial sepsis and COVID-19 [32, 33]. Elevated levels of *CCL22* have also been found in patients with IPF [34, 35]. *TNFSF14* promotes fibrosis in the cardiac muscle and atria [36], lung [37] and kidney [38]. Interestingly, *TNFSF14* has been reported to regulate fibrosis in both structural and immune cells [37] (**Fig. 4g**, **Suppl. Fig. 5c**).

IPF is driven by excessive extracellular matrix (ECM) deposition and a disrupted balance between ECM production and degradation, processes in which the matrix metalloproteinase (MMP) and tissue inhibitor of metalloproteinase (TIMP) systems, particularly in macrophages, play a cen-

tral role [39]. Among the top markers identified by our differential expression analysis were well-established profibrotic genes, including *TIMP3, MMP7*, and *MMP9*, all previously linked to IPF pathology. In addition, we uncovered a set of genes—*CCND1, CRABP2, SPON2, SPINK1, CKB*, and *MMP2*—with known roles in ECM remodeling and fibrosis-related pathways [40–45]. To further support the functional relevance of these findings, Gene Ontology enrichment analysis [46, 47] of the 16 upregulated genes in the high-attention group revealed a strong enrichment for ECM-associated processes (**Fig. 4f**).

By identifying the subset of profibrotic macrophages previously implicated in ECM remodeling and fibrotic progression, MultiMIL independently validates known pathogenic populations and suggests additional genes that may contribute to fibrotic signaling, thus expanding current disease gene signatures.

MultiMIL predicts grades in breast cancer from IMC data

MultiMIL can also be utilized directly with the raw measurements of an imaging mass cytometry (IMC) experiment and map learned attention scores to the spatial domain. We use a dataset of pathology images of breast cancer tumors [22] of different subtypes (hormone receptor (HR) +/-, HER2 +/-) and grades (1-3). The panel of 35 antibodies included breast cancer-specific biomarkers like estrogen receptor (ER), progesterone receptor (PR), HER2, the proliferation marker Ki-67, as well as lineage markers. Some of the present biomarkers were shown to be associated with higher grades, e.g., Ki67 or Vimentin [48, 49], so we focused our experiments on the grade prediction.

We trained MultiMIL to predict tumor grades (1-3) and achieved an accuracy of 0.87, substantially outperforming all baseline models, with the next-best accuracy of 0.74 from a random forest trained on cell-level input (**Fig. 5a, Suppl. Fig. 6a**). To interpret these predictions, we examined the cells with the highest attention scores and their spatial distribution. Most of these high-attention cells were tumor cells (**Fig. 5b,c**), indicating that the model relies heavily on tumor cells to distinguish between grades. Interestingly, the top-attention cells were consistently located in the tumor core rather than at the periphery. A co-occurrence analysis [50] confirmed that the model preferentially attends to tumor cells surrounded by other tumor cells (**Fig.5d**), a pattern observed across all three grades (**Suppl. Fig. 6b**).

Next, we examined how high-attention tumor cells differ from lower-attention ones. We performed differential expression testing within each grade and identified several biomarkers previously reported as dysregulated in breast cancer. For instance, GATA3 was upregulated in high-attention tumor

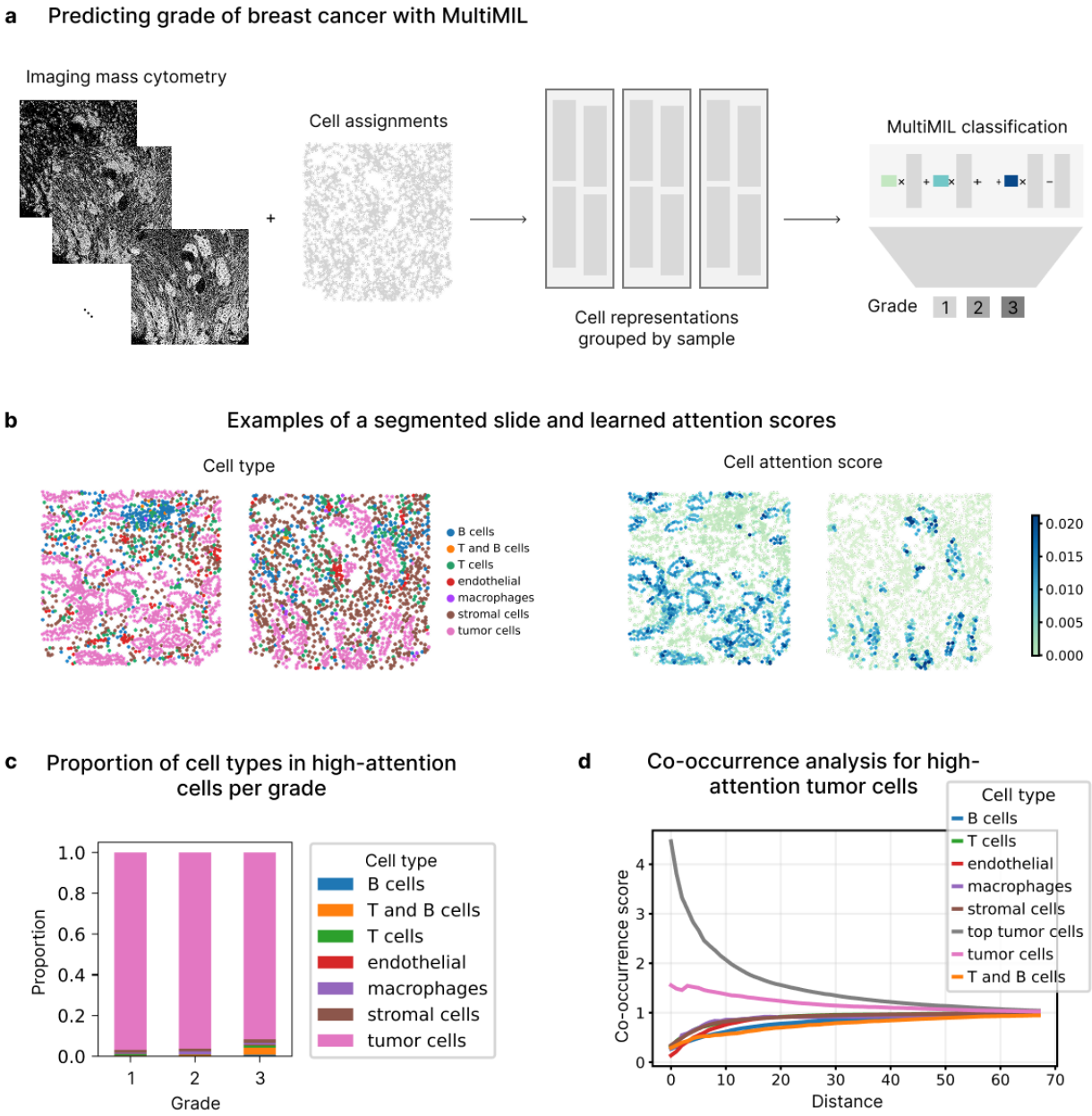


Figure 5 | MultiMIL identifies spatially co-localized grade-specific tumor cells in IMC breast cancer data. (a) MultiMIL was trained to predict grades of breast cancer from the channels of IMC data, segmented to the single-cell resolution. **(b)** Two example slides colored by cell type and attention score. **(c)** Bar plot showing the proportion of cell type among high-attention cells per grade. **(d)** Co-occurrence of high-attention tumor cells with other cell types averaged across grades, depending on the distance on the slide.

cells in grades 1 and 2 [51], while Vimentin and CD44 were enriched in grade 3 [52, 53]. To further test whether focusing on the most informative tumor cells enhances detection of grade-specific signals, we restricted the analysis to top-attention tumor cells and observed increased sensitivity: for grades 2 and 3, this approach revealed more differential biomarkers (3 and 8 vs. 4 and 11, respectively; Suppl. Table 1). Notably, p53 showed stronger differential expression between grades when restricting to top-attention tumor cells (1.5 log-fold change) compared to all tumor cells (1.26 log-fold change). Interestingly, studies have reported conflicting associations between p53 expression and tumor grade: while one study found no association between p53 and tumor grade in patients with visceral metastatic breast cancer [54], another showed that p53+ status correlated strongly with grade 3 tumors and basal subtype in African-American women [55]. This discrepancy may reflect underlying heterogeneity, where only a subset of tumor cells contributes to the observed associations. Our findings support this view and suggest that these informative subpopulations highlighted by MultiMIL may correspond to spatially co-localized tumor cells in the tumor core.

3 Discussion

MultiMIL provides a deep-learning framework for predicting sample-level phenotypes from single-cell data while simultaneously identifying disease-associated cellular states. By combining attention pooling with multiple-instance learning, our approach directly links cell-level features to sample-level outcomes, offering a scalable pipeline for atlas-scale analyses. Unlike classical aggregation strategies, MultiMIL highlights the cellular populations most predictive of disease and pinpoints the transcriptomic and compositional changes that drive these differences. Additionally, MultiMIL can be utilized to predict phenotypes for unseen samples and learn sample representations.

Multiple-instance learning has proven effective for linking information across scales. In pathology, it has been widely used to connect cellular features with whole-slide level predictions [56, 57]. Early studies have begun adapting MIL to single-cell data, either in imaging-based or genomics contexts [16, 17]. Here, we extend this idea by framing single cells as instances and biological samples as bags, demonstrating how MIL can provide a framework for analyzing single-cell genomics. Future work should systematically benchmark MIL-based approaches to identify optimal strategies for different single-cell applications.

As with other deep-learning methods, MultiMIL is prone to variability due to stochastic training and hyperparameter sensitivity. We found that attention-based cell state prioritization can fluctuate

across runs. To mitigate this, we recommend training across multiple splits and interpreting aggregated attention scores. Hence, attention weights should be viewed as hypothesis-generating signals rather than definitive markers of causal biology.

The growing availability of large-scale resources such as the Human Cell Atlas [1, 58] makes MIL-based approaches especially timely. As more large-scale atlases are released, MultiMIL can be readily applied to these datasets to identify cell states potentially relevant to various diseases. This will be particularly impactful in complex diseases such as Alzheimer's, where large cohort datasets are already available [59, 60]. MultiMIL will facilitate the discovery of novel disease-associated cell states and mechanisms.

Additionally, MultiMIL could be utilized for perturbation studies to understand how cells respond to various treatments or environmental changes. This application is crucial for identifying potential therapeutic targets and understanding drug response mechanisms [61]. By analyzing perturbation data, MultiMIL could reveal how different cell states shift in response to specific interventions, providing insights that can guide the development of patient-tailored drugs. This approach could potentially not only help identify effective treatments but also customize therapies to individual patients based on their unique cellular responses, thereby enhancing the precision and efficacy of medical interventions [62].

Several approaches for unsupervised sample representation learning have been developed in recent years [63, 64]. We envision MultiMIL as a complementary approach to such methods, allowing researchers to first identify subclusters of patients in an unsupervised manner and then utilize the clustering assignments to identify cell states driving the separation beyond known metadata. As an extension, we experimented with incorporating sample-level covariates such as sex and age by encoding them and providing them to the model. This did not lead to performance gains, suggesting that simply appending covariates may not be sufficient. Nevertheless, we see substantial potential in this direction; future research will need to investigate how covariates should be represented, where they should be introduced in the architecture, and what interactions with cellular features are most informative.

MultiMIL offers a novel approach for linking single-cell-level and sample-level data, identifying biologically meaningful disease-associated cell states. By accommodating raw data and embeddings from existing atlases or foundation models, our method provides computational biologists with a versatile and scalable tool for multi-scale analysis.

4 Conclusions

MultiMIL is a deep-learning framework that bridges the gap between single-cell measurements and sample phenotypes through attention-based multiple-instance learning. Our model learns to predict disease phenotype by prioritizing the cell states that drive the disease. This approach can be utilized with a variety of modalities and input data types, ranging from IMC data to embeddings from foundation models. We demonstrated how MultiMIL can be applied to COVID-19, idiopathic pulmonary fibrosis and breast cancer datasets, not only recovering known disease states but also expanding molecular signatures. MultiMIL is a scalable and interpretable tool for linking single-cell measurements and patient phenotypes, enabling multi-scale learning in single-cell genomics.

5 Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and materials

The package is available at http://github.com/theislab/multimil and the pipeline for sample-level prediction is available at https://github.com/theislab/sample-prediction-pipeline. The code to reproduce the results and figures is available at http://github.com/theislab/multimil_reproducibility. All datasets analyzed in this manuscript are public and can be downloaded through http://github.com/theislab/multimil_reproducibility.

Competing interests

ML owns interests in Relation Therapeutics, and is a scientific co-founder and part-time employee at AIVIVO. FJT consults for Immunai Inc., CytoReason Ltd, Cellarity, BioTuring Inc., and Genbio.AI Inc., and has an ownership interest in Dermagnostix GmbH and Cellarity.

Funding

This work was funded by the European Union (ERC, DeepCell - 101054957) and the DFG Leibniz Prize awarded to FJT.

Authors' contributions

FJT conceived the project with contributions from ML and AL. AL, ML and FJT designed the model. AL implemented the model. AL performed the benchmarks. AL, SHZ and AAM analyzed cell attention scores for the PBMC case study. AL performed the analysis for the HLCA and the breast cancer case studies. All authors contributed to the manuscript. ML and FJT supervised the project.

Acknowledgements

We thank Dr. Luke Zappia for the figure and manuscript feedback, Jan Engelmann and Vladimir Shitov for discussions on MIL and patient representation learning, Maiia Shulman for discussions on breast cancer, Dr. Daria Romanovskaia for the manuscript feedback and Dr. Malte D. Luecken for the constructive feedback throughout the project. We thank the `scverse` community (especially the developers and the maintainers of `scanpy` and `scvi-tools` packages) and the Theis lab for valuable discussions.

6 Methods

MultiMIL training

MultiMIL is a deep learning model that consists of a pooling layer and a feed-forward classification network. We assume the input to the model is single-cell embeddings, where assignments of cells to samples and sample-level labels are known. We will now focus on a single mini-batch and describe one forward pass of the model. Each training batch consists of single-cell data from l samples $\{X_1, \ldots, X_l\}$ and the sample labels $\{p_1, \ldots, p_l\}$. The number of rows in each matrix X_i equals n, which is the number of cells in the sample mini-batch, and the number of columns equals h, which is the number of features in the input space.

The single-cell representations are first fed into the pooling layer. The goal here is to aggregate the representations of all cells $z \in \mathbb{R}^{n \times h}$ from the sample (i.e., bag) into a $z_{\text{bag}} \in \mathbb{R}^h$. This bag representation corresponds to a pooled representation of a bag of cells. The pooling functions are applied independently to each latent dimension, so we will refer to the representation of cell i along latent dimension p as $z_i^p \in \mathbb{R}$ and explain how the pooled representation is calculated for one dimension. There are several ways to obtain this pooled representation, e.g., applying *max* operator

$$z_{\text{bag}}^p = \max_{i \in \text{bag}} z_i^p, \tag{1}$$

mean operator,

$$z_{\text{bag}}^p = \frac{1}{\left|\text{bag}\right|} \sum_{i \in \text{bag}} z_i^p, \tag{2}$$

where $\left|\text{bag}\right|$ indicates the number of cells in a bag. More advanced pooling functions include attention or gated attention pooling [15], where first a scalar value $a^i \in \mathbb{R}$ per cell is learned with an attention mechanism [65]:

$$a_i = \frac{\exp\left[w^T \tanh(V z_i)\right]}{\sum_{j \in \text{bag}} \exp\left[w^T \tanh(V z_j)\right]}, \tag{3}$$

or gated attention mechanism [66]:

$$a_i = \frac{\exp\left[w^T(\tanh(Vz_i) \odot \text{sigm}(Uz_i))\right]}{\sum_{j \in \text{bag}} \exp\left[w^T(\tanh(Vz_j) \odot \text{sigm}(Uz_j))\right]}. \tag{4}$$

Then the bag representation is learned as a weighted average:

$$z^p_{\text{bag}} = \sum_{i \in \text{bag}} a_i z^p_i, \tag{5}$$

where $w \in \mathbb{R}^q, V \in \mathbb{R}^{q \times h}$ and $U \in \mathbb{R}^{q \times h}$ are learnable weights and q is a hyperparameter known as attention dimension. We note that the weight a_i is learned per cell and not per cell per latent dimension.

After the aggregation, z_{bag} is fed into a classification network that consists of blocks of a linear layer with dropout, layer normalization and nonlinearity. The number of neurons in the last layer equals the number of classes for the classification tasks. The classification network predicts the distribution of disease labels for a given bag (i.e., sample). For regression tasks, the last layer consists of one neuron and predicts a scalar value per bag. Next, we will discuss the training loss.

The classification loss is calculated as the cross-entropy loss between one-hot encoded true sample labels and the predicted values of the final layer in the classification network. If the user is interested in modeling the sample labels as a progression, the last layer of the classification network can be changed to a regression head. In this case, the regression loss is calculated as mean squared error loss. To calculate the accuracy of the per-sample predictions, we round the prediction to the closest true value. For simplicity, we refer to the regression loss also as the classification loss.

MultiMIL inference

During test time, we aim to predict the disease class for new patients. We assume that the single-cell representations have the same dimensionality as the train data and that they were obtained from either mapping the query data onto the same atlas/reference with scArches [67], from the same foundation model or were left out from the train dataset as the test data. Then, we perform one forward pass through the model as described in the previous section. The pooling layer aggregates the cell representations into a bag representation, which is then classified using the classification network.

Integration metrics

To assess the quality of the integration, we used several metrics from the scIB package [23]. Note

that scIB metrics were designed for unimodal integration, and not all of them can be easily applied in the multimodal case; hence, we chose the metrics that only require the integrated embedding space as input (and not, e.g., the original unintegrated space). In the following, we briefly discuss two metrics for batch removal and four for biological variance conservation. As in scIB, the final score was calculated as 0.4*batch correction + 0.6*biological conservation. For more details on the metrics and the implementation, see [23].

Batch correction

Graph connectivity measures how well cells from each cell type are connected in a k-nearest neighbor graph. If the connectivity is high, then the batch effect was removed sufficiently. Average silhouette width (ASW) compares average distances within a cluster with distances to other clusters. The resulting score reflects how compact the clustering is. For ASW batch, we expect the batch clusters to be well-mixed together for a high batch correction score.

Biological variance conservation

Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) evaluate how well the clustering is aligned with the ground truth labels, i.e., cell type annotations. ASW label is a modification of ASW batch, where we expect the cell type clusters to be compact and separate from other cell type clusters for a high biological conservation score. Isolated label ASW assesses how well rare cell types are distinguishable from the rest of the data.

Benchmarks

Obtaining the cell embeddings

For the hyperparameter benchmark, we first obtained the scVI, PCA, GeneFormer and scGPT embeddings of the full Stephenson et al. dataset [20]. scVI was trained with the default parameters with 'Site' and 'sample_id' as batch covariates. We used helical [68] to obtain embeddings from GeneFormer (*gf-12L-95M-i4096* model) and scGPT, both in a zero-shot setting. We note that scVI embedding has 30 latent dimensions, PCA has 50, and GeneFormer and scGPT both have 512.

Benchmakring hyperparameters

We reported the accuracy and the standard deviation of MultiMIL classification and regression models on the top five performing models for each of the embeddings. The tested hyperparameters are shown in Table 1.

To assess the effects of learning rate, batch size and the pooling function, we focused on the scVI and scGPT embeddings as they scored best with respect to scIB scores. Hence, we reported across

all runs for these two embeddings.

Classification prediction

We compared MultiMIL to several baselines: random forest, multiclass logistic regression and feed-forward neural networks. We trained each model on the following data input types: mean embeddings, cell-type mean embeddings, cell-type frequency vectors and cell embeddings. We note that some baselines, namely cell-type mean embeddings and cell-type frequency vectors, require cell-type information, while MultiMIL and the rest of the baselines are entirely unsupervised.

The benchmark was performed on the Stephenson et al. PBMC dataset [20]. Throughout these experiments, we used the scVI embeddings. We created 5-fold cross-validation splits based on patient information, i.e., so that cells in each train/validation split come from different patients. We used `sklearn.model_selection.KFold()` to create the splits and `sklearn.metrics.classification_report()` to report the classification accuracy. We additionally optimized the learning rate for the IMC dataset [22] as the feature set and the measurement distribution are substantially different from those of single-cell data.

The parameters tested in the grid search are shown in Table 1.

Hyperparameter	Description	Default	Range
Learning rate	learning rate parameter	1e-4 for single-cell, 1e-3 for IMC	$\{1e-5, 1e-4, 1e-3, 1e-2\}$
Batch size	number of cells in a forward pass	256	$\{256, 512, 1024\}$
Pooling function	how the sample representations are calculated	gated attention	{gated attention, attention, mean, max}
Epochs	number of epochs the model was trained for	50	{50, 100, 150, 200}

Table 1 | Hyperparameter search for MultiMIL's prediction.

Depending on the batch size, the number of samples in each batch varied from 2 to 8, as 128 cells from each sample were drawn for the forward pass. The default parameters were chosen based on the prediction accuracy of the validation set averaged across five splits.

Next, we discuss the baseline models and the input data in more detail. We performed a hyperparameter grid search for NN-based models and reported the best-performing configuration. Patient disease labels were used as class labels throughout this benchmark apart from the "Cell embedding" input type, where all the cells from a diseased donor were assumed to have the disease class.

Baseline models

- Multiclass logistic regression is an extension of the logistic regression method that allows the prediction of multiple classes. We calculate the probability of belonging to a particular class with a softmax function and calculate the loss as the entropy between predicted probabilities and the true class. We optimize the loss function with gradient descent.

- Random forest was implemented using `sklearn.ensemble.RandomForestClassifier()` with the default parameters.
- Neural network was implemented as a 2-layer feed-forward network with one hidden layer of 64 neurons, batch normalization and ReLU activation. The second linear layer outputs class probabilities. We trained the neural network baselines with Adam optimizer [69] for 200 epochs for sample-level inputs and 30 epochs for cell-level inputs. Hyperparameter search was run for batch size and learning rate shown in Table 2.

Hyperparameter	Description	Range
Learning rate	learning rate parameter	$\{1e-5, 1e-4, 1e-3\}$
Batch size for sample-level inputs	size of the training mini-batch	{8, 16, 32, 64}
Batch size for cell-level input	size of the training mini-batch	{128, 256, 512, 1024}

Table 2 | Hyperparameter search NN baseline.

Input data types

- Mean embedding representations were calculated from the latent embeddings with `decoupler.get_pseudobulk()` specifying the sample parameter and keeping all the cells.
- Cell type-aware mean embedding representations were calculated from the latent embeddings with `decoupler.get_pseudobulk()` specifying the sample and group (i.e., cell type) parameters and keeping all the cells. To obtain one representation per sample, we concatenated cell-type-specific vectors into one vector.
- Frequency vectors were calculated from cell type proportions for each sample.
- Low-dimensional cell embeddings were directly passed to the baselines, e.g., from atlases or foundation models.

Identification of DA cell states

We ran the default Milo [13] analysis on the PBMC dataset using the embeddings from scVI following the pertpy [70] implementation. We ran three pairwise analyses comparing mild COVID-19 and healthy, severe COVID-19 and healthy, and severe and mild COVID-19. We show the neighborhoods with spatial false discovery rate (FDR) corrected levels of less than 0.05.

To identify cell types with the highest compositional change between healthy and severe (**Fig. 3**e), we excluded the cell types with fewer than 1,000 cells and calculated the relative change in proportions between the number of cells in pooled severe and healthy samples.

Sample representations

We also investigated how well we can predict sample labels with a kNN classifier from sample representations obtained by aggregating cell embeddings with different aggregation strategies. We used the attention scores from the best-performing model trained on the scVI embedding throughout these comparisons. Sample representations were calculated as either the mean of cell embeddings belonging to the sample or the weighted average (where the weight for each cell was the learned attention score). We used cells with top 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90% highest attention score for the aggregation.

Calculation of the profibrotic signature

To calculate the profibrotic score for macrophages in HLCA, we used the signature from [29]: *SPP1, LIPA, LPL, FDX1, SPARC, MATK, GPC4, PALLD, MMP7, MMP9, CHIT1, CSTK, CHI3L1, CSF1, FCMR, TIMP3, COL22A1, SIGLEC15, CCL2*. The score was calculated with `scanpy.tl.score_genes()`. We performed a two-sided t-test to check for the significance of the score in all IPF macrophages vs. IPF macrophages with the high attention score using `scipy.stats.ttest_ind()`. We used edgeR-QLF [12] to identify the genes differentially expressed in IPF macrophages with high attention compared to all IPF macrophages and reported genes with a log-fold change greater than 1.5 and FDR-corrected p-value less than 0.01 as up-regulated (see Supplementary Table 1).

Gene Ontology analysis

We used GOATOOLS [71] to run the GO term analysis on the genes that were identified as significantly upregulated in the IPF macrophages with MultiMIL. We followed the tutorial and ran all the functions with their default parameters. We reported the terms with the corrected p-value less than 0.1 as significant.

Identification of differential signal in IMC data

We performed a Wilcoxon rank-sum test between the high-attention (top 10%) tumor cells, comparing cells from each of the three grades with the rest of the high-attention tumor cells, and reported the Benjamini-Hochberg corrected p-values.

Datasets

<u>Stephenson et al.</u>

The PBMC dataset contains 647,366 cells from 130 donors, collected at three sites. All scRNA-seq data points were used for the integration. For the prediction experiment with all COVID-19 stages, we removed non-COVID and non-healthy samples, so the prediction was performed for 6 classes:

healthy (23 samples), asymptomatic (9), mild (23), moderate (30), severe (13) and critical (15). For the 3-class prediction, we subset the data to 13 healthy, 13 mild and 13 severe samples to perform the benchmarking experiments on the balanced data and not confound the performance with unbalanced classes. For the binary experiment, the COVID-19 class had 90 samples and the healthy class had 23 samples. We used 'sex' and 'age interval' as categorical sample covariates in the experiments where we concatenated one-hot covariates to the sample embeddings before passing them through the classification network.

Sikkema et al.

Human Lung Cell Atlas (HLCA) consists of the core (584,444 cells, 107 donors) and the extension datasets (1,797,714 cells, 380 donors). The core samples are all healthy, while the extension has healthy and diseased samples. In our experiments, we subset the data to healthy and IPF samples in a balanced way, i.e., the number of donors is the same (67) in both groups.

Jackson et al.

The human breast cancer dataset consists of two cohorts, the Zurich and the Basel cohorts. In our experiments, we subsetted the data to the Zurich cohort and excluded metastatic samples from the analysis. This dataset consists of 72 donors, 347 samples and 395,769 cells. We used 'core' as the sample covariate in our experiments. We trained the model to predict grades 1 (95), 2 (122) and 3 (130).

Running time

We provide training times for MultiMIL with the default architecture in Table 3. The training was performed on the same GPU server with the following characteristics: Intel(R) Xeon(R) Platinum 8280L CPU with 28 cores, 2.70GHz, Tesla V100-SXM3-32GB GPU. We report the average run time and standard deviation across three runs. We used the PBMC CITE-seq dataset [20], subsetted to healthy, mild and severe COVID-19 in a balanced way, resulting in 256,051 cells. All models were trained for 50 epochs. We modeled the prediction task as either a three-class classification problem or a regression problem.

	average runtime (s)	standard deviation (s)
classification setting	356	9
regression seting	357	5

Table 3 | Running times for MultiMIL.

Default architecture

The default architecture for the classification setting of MultiMIL is shown below. In the regression setting, the last layer has only one output neuron.

	Layer
MIL attention pooling	calculation of attention scores as in Eq. 4
	calculation of the weighted sum as in Eq. 5
Classifier network	Linear(embed_dim, 128)
	Dropout(0.2)
	LayerNorm
	LeakyReLU
	Linear(128, n_classes)

Table 4 | Default architecture for MultiMIL.

We note that embed_dim refers to the number of latent dimensions in the low-dimensional representation of single-cell data from atlases or foundation models, or the raw dimensionality of the input if raw data is used. For the model versions, where we incorporated one-hot encoded sample covariates, the embed_dim equals the number of latent dimensions in the low-dimensional representation of single-cell data plus the number of categories for each of the covariates. For instance, if we concatenate the sex covariate with two categories present, one-hot encoded embeddings will add two dimensions to embed_dim.

References

[1] Hrovatin K, Sikkema L, Shitov VA, Heimberg G, Shulman M, Oliver AJ, et al. Considerations for building and using integrated single-cell atlases. Nat Methods. 2025 Jan;22(1):41-57.

[2] Svensson V, Vento-Tormo R, Teichmann SA. Exponential scaling of single-cell RNA-seq in the last decade. arXiv [q-bioGN]. 2017 Apr.

[3] Mao Y, Lin YY, Wong NKY, Volik S, Sar F, Collins C, et al. Phenotype prediction from single-cell RNA-seq data using attention-based neural networks. Bioinformatics. 2024 Feb;40(2).

[4] Ravindra N, Sehanobish A, Pappalardo JL, Hafler DA, van Dijk D. Disease state prediction from single-cell data using graph attention networks. In: Proceedings of the ACM Conference on Health, Inference, and Learning. CHIL '20. New York, NY, USA: Association for Computing Machinery; 2020. p. 121-30.

[5] Dann E, Cujba AM, Oliver AJ, Meyer KB, Teichmann SA, Marioni JC. Precise identification of cell states altered in disease using healthy single-cell references. Nat Genet. 2023 Oct.

[6] Zeng F, Kong X, Yang F, Chen T, Han J. scPheno: A deep generative model to integrate scRNA-seq with disease phenotypes and its application on prediction of COVID-19 pneumonia and severe assessment; 2022.

[7] Bruggner RV, Bodenmiller B, Dill DL, Tibshirani RJ, Nolan GP. Automated identification of stratifying signatures in cellular subpopulations. Proc Natl Acad Sci U S A. 2014 Jul;111(26):E2770-7.

[8] De Donno C, Hediyeh-Zadeh S, Moinfar AA, Wagenstetter M, Zappia L, Lotfollahi M, et al. Population-level integration of single-cell datasets enables multi-scale analysis across samples. Nat Methods. 2023 Nov;20(11):1683-92.

[9] Boyeau P, Hong J, Gayoso A, Jordan MI, Azizi E, Yosef N. Deep generative modeling for quantifying sample-level heterogeneity in single-cell omics; 2022.

[10] Xiong G, Bekiranov S, Zhang A. ProtoCell4P: An Explainable Prototype-based Neural Network for Patient Classification Using Single-cell RNA-seq. Bioinformatics. 2023 Aug.

[11] Skinnider MA, Squair JW, Kathe C, Anderson MA, Gautier M, Matson KJE, et al. Cell type prioritization in single-cell data. Nat Biotechnol. 2021 Jan;39(1):30-4.

[12] Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010 Jan;26(1):139-40.

[13] Dann E, Henderson NC, Teichmann SA, Morgan MD, Marioni JC. Differential abundance testing on single-cell data using k-nearest neighbor graphs. Nat Biotechnol. 2022 Feb;40(2):245-53.

[14] Liu T, De Brouwer E, Kuo T, Diamant N, Missarova A, Wang H, et al. Learning multicellular representations of single-cell transcriptomics data enables characterization of patient-level disease states. Bioinformatics. 2024 Nov;(biorxiv;2024.11.18.624166v1).

[15] Ilse M, Tomczak JM, Welling M. Attention-based Deep Multiple Instance Learning. arXiv; 2018. Available from: https://arxiv.org/abs/1802.04712.

[16] Sadafi A, Makhro A, Bogdanova A, Navab N, Peng T, Albarqouni S, et al.. Attention based Multiple Instance Learning for Classification of Blood Cell Disorders; 2020. Available from: https://arxiv.org/abs/2007.11641.

[17] Engelmann JP, Palma A, Tomczak JM, Theis FJ, Casale FP. Mixed Models with Multiple Instance Learning; 2024. Available from: https://arxiv.org/abs/2311.02455.

[18] Kim Y, Wang T, Xiong D, Wang X, Park S. Multiple instance neural networks based on sparse attention for cancer detection using T-cell receptor sequences. BMC Bioinformatics. 2022 Nov;23(1):469.

[19] Do C, Lähdesmäki H. Incorporating hierarchical information into multiple instance learning for patient phenotype prediction with single-cell RNA-sequencing data. Bioinformatics. 2025 Jul;41(Supplement_1):i96-i104.

[20] Stephenson E, Reynolds G, Botting RA, Calero-Nieto FJ, Morgan MD, Tuong ZK, et al. Single-cell multi-omics analysis of the immune response in COVID-19. Nat Med. 2021 May;27(5):904-16.

[21] Sikkema L, Ramírez-Suástegui C, Strobl DC, Gillett TE, Zappia L, Madissoon E, et al. An integrated cell atlas of the lung in health and disease. Nat Med. 2023 Jun;29(6):1563-77.

[22] Jackson HW, Fischer JR, Zanotelli VRT, Ali HR, Mechera R, Soysal SD, et al. The single-cell pathology landscape of breast cancer. Nature. 2020 Feb;578(7796):615-20.

[23] Luecken MD, Büttner M, Chaichoompu K, Danese A, Interlandi M, Mueller MF, et al. Benchmarking atlas-level data integration in single-cell genomics. Nat Methods. 2022 Jan;19(1):41-50.

[24] Sun D, Guan X, Moran AE, Wu LY, Qian DZ, Schedin P, et al. Identifying phenotype-associated subpopulations by integrating bulk and single-cell sequencing data. Nat Biotechnol. 2022 Apr;40(4):527-38.

[25] Guo C, Li B, Ma H, Wang X, Cai P, Yu Q, et al. Single-cell analysis of two severe COVID-19 patients reveals a monocyte-associated and tocilizumab-responding cytokine storm. Nat Commun. 2020 Aug;11(1):3924.

[26] Silvin A, Chapuis N, Dunsmore G, Goubet AG, Dubuisson A, Derosa L, et al. Elevated calprotectin and abnormal myeloid cell subsets discriminate severe from mild COVID-19. Cell. 2020 Sep;182(6):1401-18.e18.

[27] Rood JE, Maartens A, Hupalowska A, Teichmann SA, Regev A. Impact of the Human Cell Atlas on medicine. Nat Med. 2022 Dec;28(12):2486-96.

[28] Morse C, Tabib T, Sembrat J, Buschur KL, Bittar HT, Valenzi E, et al. Proliferating SPP1/MERTK-expressing macrophages in idiopathic pulmonary fibrosis. Eur Respir J. 2019 Aug;54(2).

[29] Adams TS, Schupp JC, Poli S, Ayaub EA, Neumark N, Ahangari F, et al. Single-cell RNA-seq reveals ectopic and aberrant lung-resident cell populations in idiopathic pulmonary fibrosis. Sci Adv. 2020 Jul;6(28):eaba1983.

[30] Reyfman PA, Walter JM, Joshi N, Anekalla KR, McQuattie-Pimentel AC, Chiu S, et al. Single-cell transcriptomic analysis of human lung provides insights into the pathobiology of pulmonary fibrosis. Am J Respir Crit Care Med. 2019 Jun;199(12):1517-36.

[31] Jaeger B, Schupp JC, Plappert L, Terwolbeck O, Artysh N, Kayser G, et al. Airway basal cells show a dedifferentiated KRT17highPhenotype and promote fibrosis in idiopathic pulmonary fibrosis. Nat Commun. 2022 Sep;13(1):5637.

[32] Wu Y, Wang Q, Li M, Lao J, Tang H, Ming S, et al. SLAMF7 regulates the inflammatory response in macrophages during polymicrobial sepsis. J Clin Invest. 2023 Mar;133(6).

[33] Simmons DP, Nguyen HN, Gomez-Rivas E, Jeong Y, Jonsson AH, Chen AF, et al. SLAMF7

engagement superactivates macrophages in acute and chronic inflammation. Sci Immunol. 2022 Feb;7(68):eabf2846.

[34] Inoue T, Fujishima S, Ikeda E, Yoshie O, Tsukamoto N, Aiso S, et al. CCL22 and CCL17 in rat radiation pneumonitis and in human idiopathic pulmonary fibrosis. Eur Respir J. 2004 Jul;24(1):49-56.

[35] Yogo Y, Fujishima S, Inoue T, Saito F, Shiomi T, Yamaguchi K, et al. Macrophage derived chemokine (CCL22), thymus and activation-regulated chemokine (CCL17), and CCR4 in idiopathic pulmonary fibrosis. Respir Res. 2009 Aug;10(1):80.

[36] Wu Y, Zhan S, Chen L, Sun M, Li M, Mou X, et al. TNFSF14/LIGHT promotes cardiac fibrosis and atrial fibrillation vulnerability via PI3Kγ/SGK1 pathway-dependent M2 macrophage polarisation. J Transl Med. 2023 Aug;21(1):544.

[37] Herro R, Croft M. The control of tissue fibrosis by the inflammatory molecule LIGHT (TNF Superfamily member 14). Pharmacol Res. 2016 Feb;104:151-5.

[38] Li Y, Tang M, Han B, Wu S, Li SJ, He QH, et al. Tumor necrosis factor superfamily 14 is critical for the development of renal fibrosis. Aging. 2020 Nov;12(24):25469-86.

[39] Zhao X, Chen J, Sun H, Zhang Y, Zou D. New insights into fibrosis from the ECM degradation perspective: the macrophage-MMP-ECM interaction. Cell Biosci. 2022 Jul;12(1):117.

[40] Keophiphath M, Achard V, Henegar C, Rouault C, Clément K, Lacasa D. Macrophage-secreted factors promote a profibrotic phenotype in human preadipocytes. Mol Endocrinol. 2009 Jan;23(1):11-24.

[41] Zeng S, Chen XI, Yi Q, Thakur A, Yang H, Yan Y, et al. CRABP2 regulates infiltration of cancer-associated fibroblasts and immune response in melanoma. Oncol Res. 2023 Dec;32(2):261-72.

[42] Zhang J, Liu G, Liu Y, Yang P, Xie J, Wei X. The biological functions and related signaling pathways of SPON2. Front Oncol. 2023;13:1323744.

[43] Xu L, Lu C, Huang Y, Zhou J, Wang X, Liu C, et al. SPINK1 promotes cell growth and metastasis of lung adenocarcinoma and acts as a novel prognostic biomarker. BMB Rep. 2018 Dec;51(12):648-53.

[44] Loo JM, Scherl A, Nguyen A, Man FY, Weinberg E, Zeng Z, et al. Extracellular metabolic energetics can promote cancer progression. Cell. 2015 Jan;160(3):393-406.

[45] Stamenkovic I. Extracellular matrix remodelling: the role of matrix metalloproteinases. J Pathol. 2003 Jul;200(4):448-64.

[46] Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000 May;25(1):25-9.

[47] Gene Ontology Consortium, Aleksander SA, Balhoff J, Carbon S, Cherry JM, Drabkin HJ, et al. The Gene Ontology knowledgebase in 2023. Genetics. 2023 May;224(1).

[48] Inwald EC, Klinkhammer-Schalke M, Hofstädter F, Zeman F, Koller M, Gerstenhauer M, et al. Ki-67 is a prognostic parameter in breast cancer patients: results of a large population-based cohort of a cancer registry. Breast Cancer Res Treat. 2013 Jun;139(2):539-52.

[49] Domagala W, Woźniak L, Lasota J, Weber K, Osborn M. Vimentin is preferentially expressed in high-grade ductal and medullary, but not in lobular breast carcinomas. Am J Pathol. 1990 Nov;137(5):1059-64.

[50] Palla G, Spitzer H, Klein M, Fischer D, Schaar AC, Kuemmerle LB, et al. Squidpy: a scalable framework for spatial omics analysis. Nat Methods. 2022 Feb;19(2):171-8.

[51] Gonzalez RS, Wang J, Kraus T, Sullivan H, Adams AL, Cohen C. GATA-3 expression in male and female breast cancers: comparison of clinicopathologic parameters and prognostic relevance. Hum Pathol. 2013 Jun;44(6):1065-70.

[52] McFarlane S, Coulter JA, Tibbits P, O'Grady A, McFarlane C, Montgomery N, et al. CD44 increases the efficiency of distant metastasis of breast cancer. Oncotarget. 2015 May;6(13):11465-76.

[53] Domagala W, Lasota J, Bartkowiak J, Weber K, Osborn M. Vimentin is preferentially expressed in human breast carcinomas with low estrogen receptor and high Ki-67 growth fraction. Am J Pathol. 1990 Jan;136(1):219-27.

[54] Yang P, Du CW, Kwan M, Liang SX, Zhang GJ. The impact of p53 in predicting clinical outcome of breast cancer patients with visceral metastasis. Sci Rep. 2013 Jul;3(1):2246.

[55] Dookeran KA, Dignam JJ, Ferrer K, Sekosan M, McCaskill-Stevens W, Gehlert S. p53 as a marker of prognosis in African-American women with breast cancer. Ann Surg Oncol. 2010 May;17(5):1398-405.

[56] Li B, Li Y, Eliceiri KW. Dual-stream Multiple Instance Learning Network for Whole Slide Image Classification with Self-supervised Contrastive Learning. Conf Comput Vis Pattern Recognit Workshops. 2021 Jun;2021:14318-28.

[57] Javed SA, Juyal D, Padigela H, Taylor-Weiner A, Yu L, Prakash A. Additive MIL: Intrinsically Interpretable Multiple Instance Learning for Pathology; 2022. Available from: https://arxiv.org/abs/2206.01794.

[58] Regev A, Teichmann SA, Lander ES, Amit I, Benoist C, Birney E, et al. The Human Cell Atlas. Elife. 2017 Dec;6.

[59] Mathys H, Peng Z, Boix CA, Victor MB, Leary N, Babu S, et al. Single-cell atlas reveals correlates of high cognitive function, dementia, and resilience to Alzheimer's disease pathology. Cell. 2023 Sep;186(20):4365-85.e27.

[60] Mathys H, Davila-Velderrain J, Peng Z, Gao F, Mohammadi S, Young JZ, et al. Single-cell transcriptomic analysis of Alzheimer's disease. Nature. 2019 Jun;570(7761):332-7.

[61] Ji Y, Lotfollahi M, Wolf FA, Theis FJ. Machine learning for perturbational single-cell omics. Cell Syst. 2021 Jun;12(6):522-37.

[62] Gavriilidis GI, Vasileiou V, Orfanou A, Ishaque N, Psomopoulos F. A mini-review on perturbation modelling across single-cell omic modalities. Comput Struct Biotechnol J. 2024 Dec;23:1886-96.

[63] Wang H, Torous W, Gong B, Purdom E. Visualizing scRNA-Seq data at population scale with GloScope. Genome Biol. 2024 Oct;25(1):259.

[64] Joodaki M, Shaigan M, Parra V, Bülow RD, Kuppe C, Hölscher DL, et al. Detection of PatIent-Level distances from single cell genomics and pathomics data with Optimal Transport (PILOT). Mol Syst Biol. 2024 Feb;20(2):57-74.

[65] Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate; 2016. Available from: https://arxiv.org/abs/1409.0473.

[66] Dauphin YN, Fan A, Auli M, Grangier D. Language Modeling with Gated Convolutional Networks; 2017.

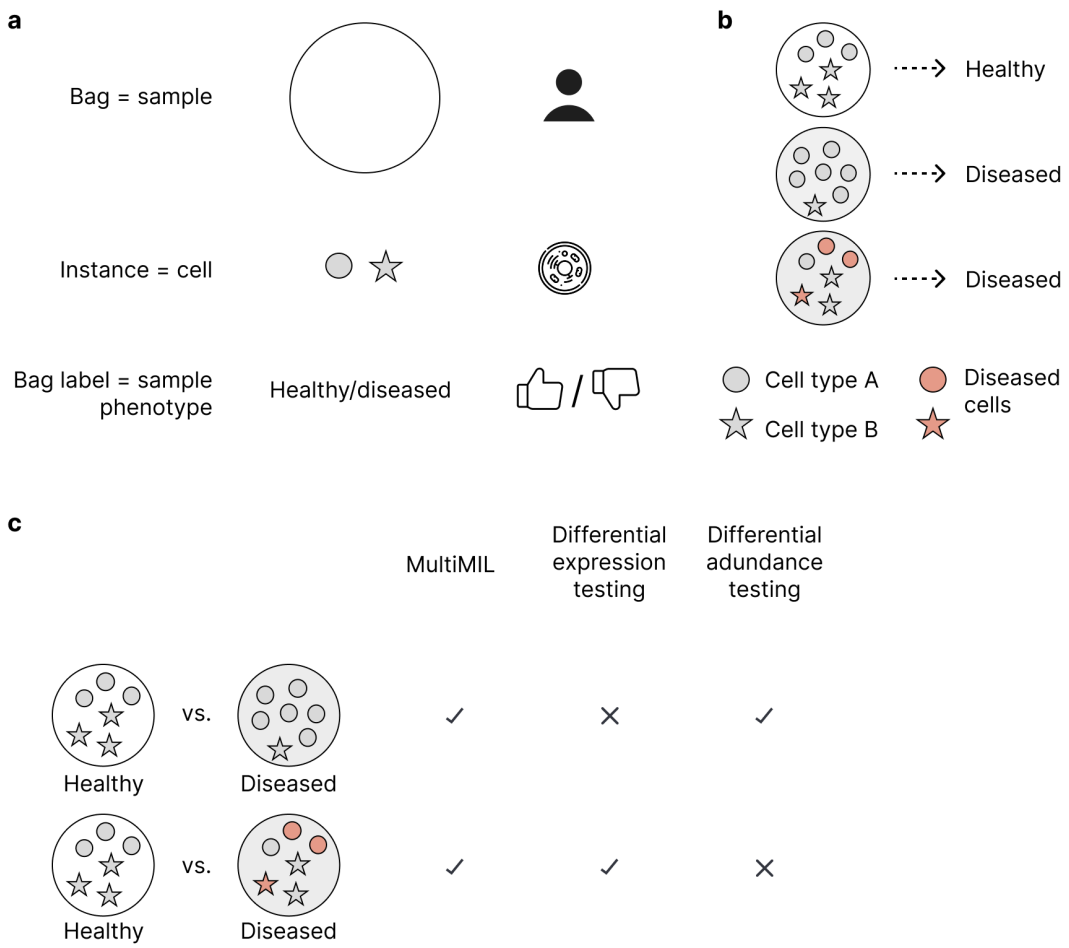
[67] Lotfollahi M, Naghipourfar M, Luecken MD, Khajavi M, Büttner M, Wagenstetter M, et al. Mapping single-cell data to reference atlases by transfer learning. Nat Biotechnol. 2022 Jan;40(1):121-30.

[68] Team H. helicalAI/helical: v1.1.0. Zenodo; 2024. Available from: https://doi.org/10.5281/zenodo.13135902.

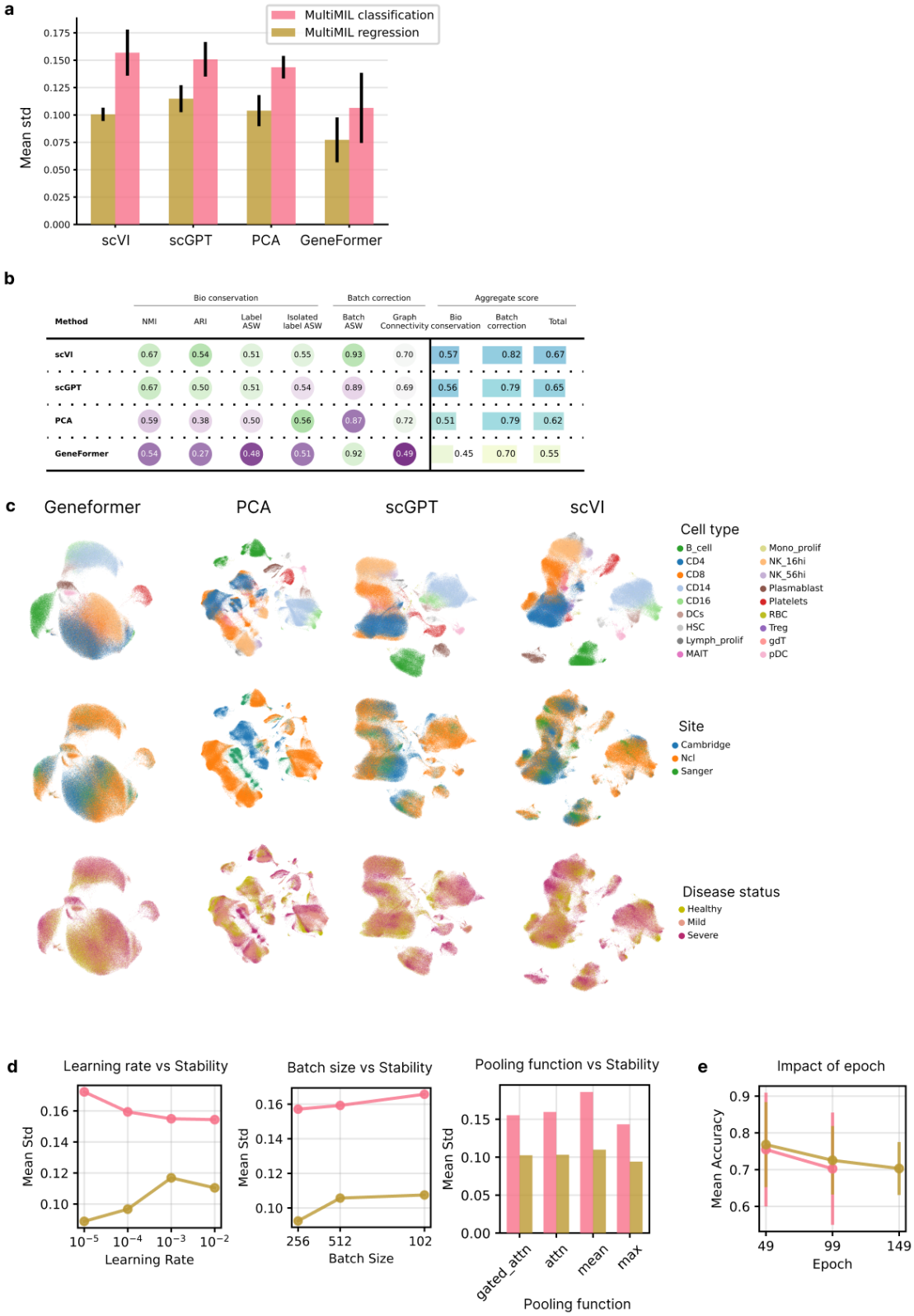
[69] Kingma DP, Ba J. Adam: A Method for Stochastic Optimization; 2017. Available from: https://arxiv.org/abs/1412.6980.

[70] Heumos L, Ji Y, May L, Green T, Zhang X, Wu X, et al. Pertpy: an end-to-end framework for perturbation analysis. bioRxiv. 2024. Available from: https://www.biorxiv.org/content/early/2024/08/07/2024.08.04.606516.

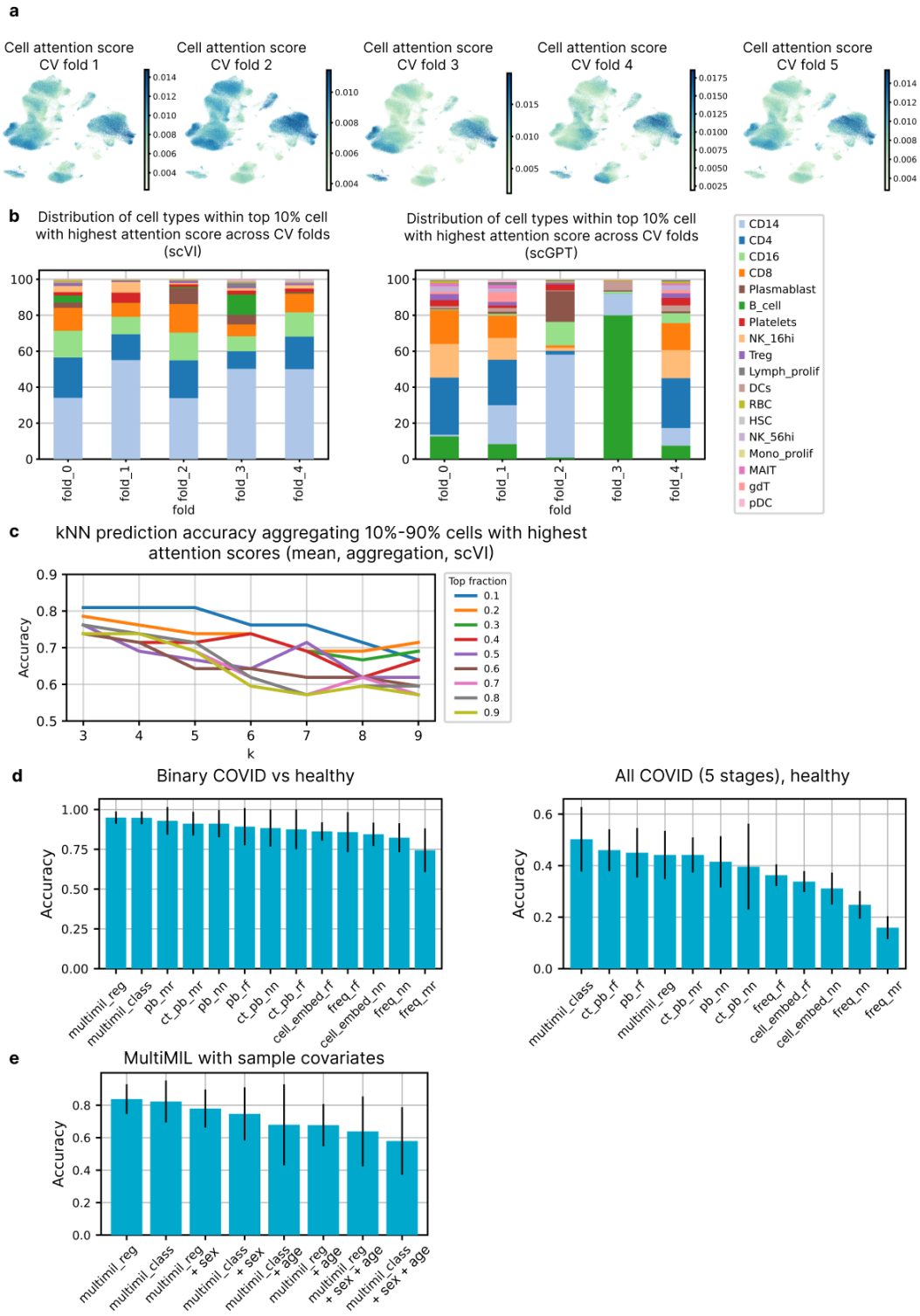
[71] Klopfenstein DV, Zhang L, Pedersen BS, Ramírez F, Warwick Vesztrocy A, Naldi A, et al. GOATOOLS: A Python library for Gene Ontology analyses. Sci Rep. 2018 Jul;8(1):10872.



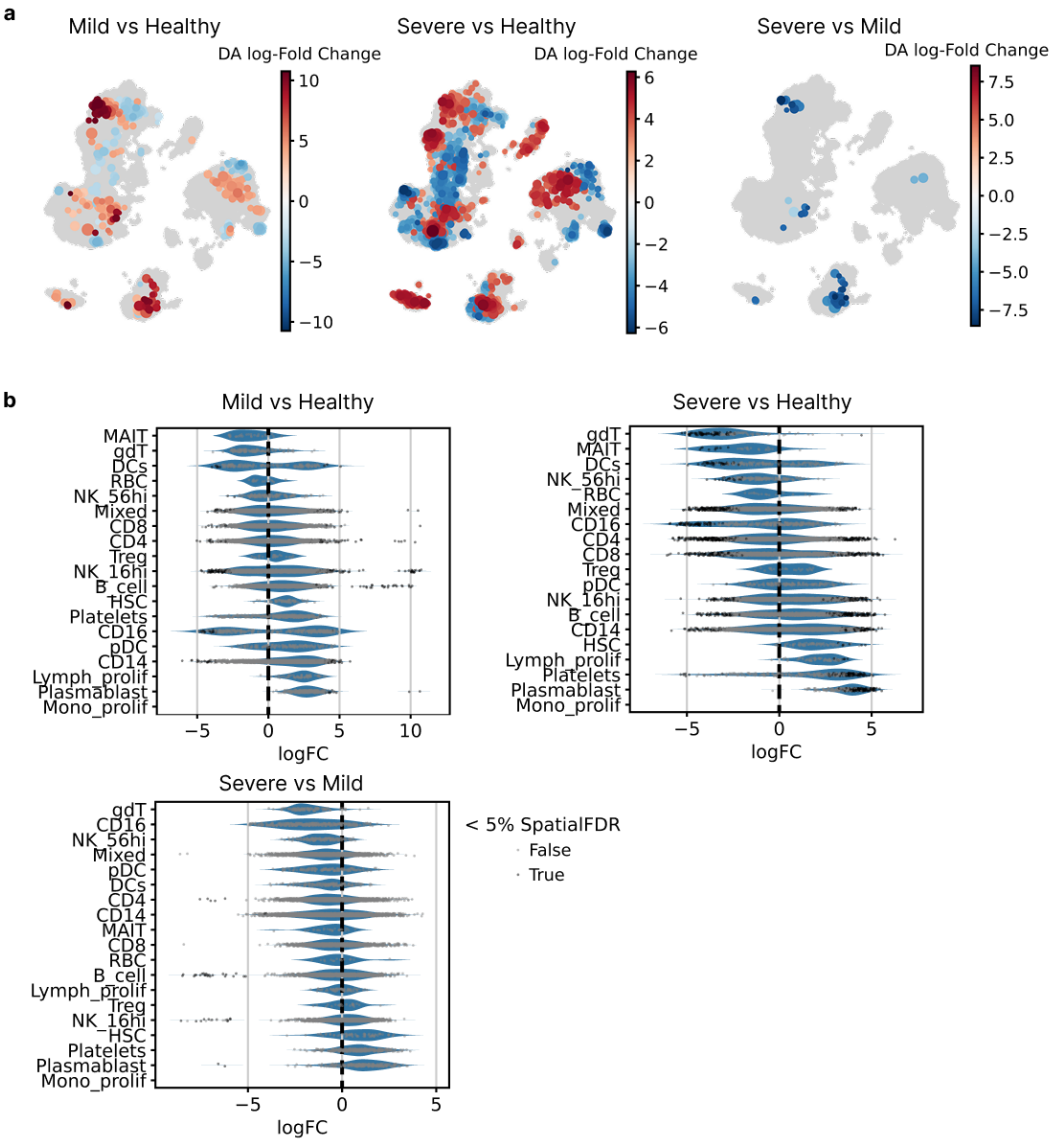
Supplementary Figure 1 | Multiple instance learning. **(a)** In our context, bags correspond to samples, instances to cells and the classification labels are known for bags. **(b)** Examples of data points in the multiple-instance-learning dataset. Our task is to classify bags into classes and identify cells (i.e. colored instances) that are associated with a certain disease. **(c)** MultiMIL can identify changes in the abundance of cell types between conditions as well as transcriptomic changes.



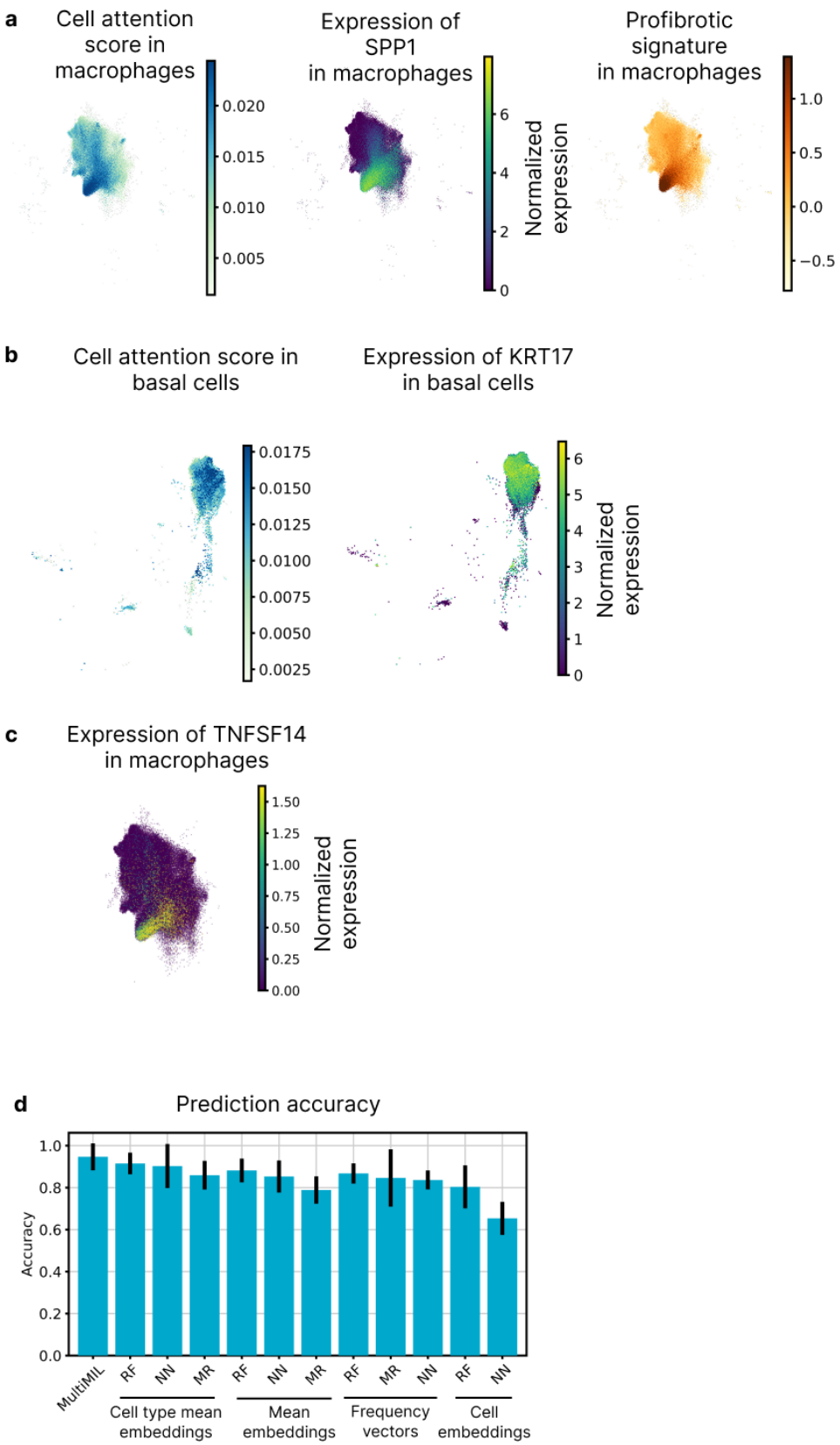
Supplementary Figure 2 | Benchmarking MultiMIL on different embeddings for Stephenson et al. **(a)** A bar plot showing the mean standard deviation of prediction accuracy across cross-validation splits for scVI, PCA, scGPT and GeneFormer embeddings, taking the top 5 MultiMIL models, colored by regression/classification setting. **(b)** scIB metrics assessing the quality of the embeddings. **(c)** UMAPs of the GeneFormer, PCA, scGPT and scVI embeddings, colored by cell type, batch (i.e., site) and disease stage. **(d)** Plots showing the mean standard deviation of prediction accuracy depending on the learning rate, batch size and the pooling method, colored by MultiMIL's regression/classification setting. **(e)** Effect of the epoch, i.e., how long the model was trained for, on the accuracy.



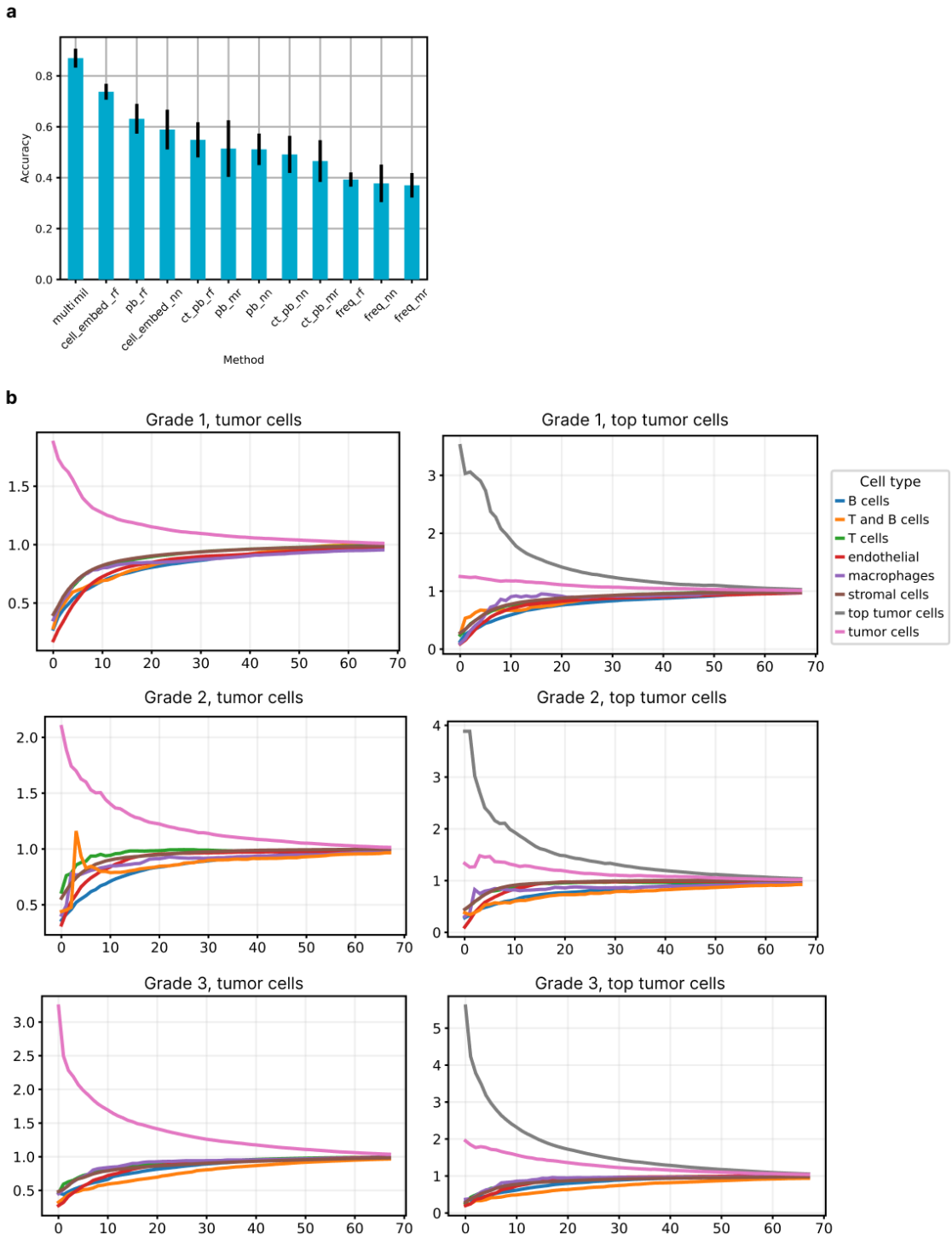
Supplementary Figure 3 | MultiMIL's cell attention scores, sample representations and prediction accuracy across different tasks. **(a)** UMAPs showing cell attention scores learned in five cross-validation runs using scVI embedding and MultiMIL's regression setting. **(b)** Stacked bar plots showing the distribution of cell types with top 10% highest attention scores across five cross-validation runs, using top-performing MultiMIL's regression models on scVI and scGPT embeddings. **(c)** Line plots showing how well the kNN classifier can predict sample labels from $3, 4, 5, 6, 7, 8, 9$ nearest neighbors when the sample representation was obtained by averaging cell embeddings of cells with top 10%-90% highest attention scores. **(d)** Results of the prediction benchmark on binary (healthy, COVID-19) and full data (healthy, 5 COVID-19 stages) using scVI embeddings, comparing MultiMIL with the baselines. **(e)** Results of the prediction benchmark on the three-class (healthy, mild and severe COVID-19) task using scVI embeddings, comparing MultiMIL default models with MultiMIL models incorporating sample covariates (sex and age).



Supplementary Figure 4 | MILO analysis of Stephenson et al. dataset. (a) Results of Milo analysis run on MultiMIL's embeddings, mild vs. healthy, severe vs. healthy and severe vs. mild, each colored by DA log-fold change (red corresponds to the first condition in the title). **(b)** Violin plots showing DA changes for each of the cell types in mild vs. healthy, severe vs. healthy and severe vs. mild.



Supplementary Figure 5 | IPF in HLCA. **(a)** UMAPs of macrophages, colored by cell attention, normalized expression of SSP1 and profibrotic signature score. **(b)** UMAPs of basal cells, colored by cell attention and normalized expression of KRT17. **(c)** The UMAP of macrophages, colored by the normalized expression of TNFSF14. **(d)** Prediction accuracy of MultiMIL compared to other baseline models on the binary classification task, i.e., IPF vs healthy.



Supplementary Figure 6 | Predicting grades in breast cancer from IMC data. **(a)** Prediction accuracy of MultiMIL compared with baselines. **(b)** Co-occurrence analysis of tumor cells (left columns) and high-attention tumor cells (i.e., top tumor cells), split by grade.