Journal Title Here, 2022, pp. 1–13

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

scMILD: Single-cell Multiple Instance Learning for Sample Classification and Associated Subpopulation Discovery

Kyeonghun Jeong,[1] Jinwook Choi[1,2*] and Kwangsoo Kim[3,4*]

[1]Interdisciplinary Program in Bioengineering, Seoul National University, 1 Gwanak-ro, 08826, Seoul, Republic of Korea, [2]Institute of Medical and Biological Engineering, Medical Research Center, Seoul National University, 103 Daehak-ro, 03080, Seoul, Republic of Korea, [3]Department of Transdisciplinary Medicine, Institute of Convergence Medicine with Innovative Technology, Seoul National University Hospital, 101 Daehak-ro, 03080, Seoul, Republic of Korea and [4]Department of Medicine, College of Medicine, Seoul National University, 101 Daehak-ro, 03080, Seoul, Republic of Korea
*Corresponding author. jinchoi@snu.ac.kr, kwangsookim@snu.ac.kr
FOR PUBLISHER ONLY Received on Date Month Year; revised on Date Month Year; accepted on Date Month Year

Abstract

Single-cell transcriptomics enables the study of cellular heterogeneity, but current unsupervised strategies make it challenging to associate individual cells with sample conditions. We propose scMILD, a weakly supervised learning framework based on Multiple Instance Learning, which leverages sample-level labels to identify condition-associated cell subpopulations. scMILD employs a dual-branch architecture to perform sample-level classification and cell-level representation learning simultaneously. We validated the model's reliable identification of condition-associated cells using controlled simulation studies with CRISPR-perturbed cells. Evaluated on diverse single-cell RNA-seq datasets, including Lupus, COVID-19, and Ulcerative Colitis, scMILD consistently outperformed state-of-the-art models and identified condition-specific cell subpopulations consistent with the original studies' findings. This demonstrates scMILD's potential for exploring cellular heterogeneity underlying various biological conditions and its applicability in different disease contexts.

Key words: Single-cell transcriptomics, Multiple Instance Learning, Sample classification, Condition-associated cell subpopulations

Key Messages

- scMILD: A novel weakly supervised framework for single-cell transcriptomics
- Dual-branch architecture enables sample classification and cell subpopulation identification
- Outperforms state-of-the-art models across diverse single-cell RNA-seq datasets
- Identifies biologically relevant condition-associated cell subpopulations
- Bridges the gap between sample-level phenotypes and cellular heterogeneity

Introduction

Single-cell transcriptomic data derived from various samples with diverse conditions and phenotypes have been accumulating rapidly [14, 5, 10]. However, studies directly analyzing the association between sample conditions/phenotypes and individual cells still need to be completed. The primary analysis of single-cell RNA-seq often focuses on identifying condition/phenotype-specific or associated cell subsets [17, 24, 16]. This process typically relies on unsupervised strategies, which can be time-consuming, require extensive knowledge of conditions/phenotypes, and involve subjective judgment from the analyst. These approaches also present challenges for reproducibility and validation, making it difficult to associate individual cells with sample conditions directly.

Recent computational advances have attempted to address these challenges. Deep learning approaches have enabled reference mapping with interpretable gene programs while improving clustering accuracy through various architectural innovations [6]. Several methods have focused on improving cell type identification and clustering accuracy through novel deep learning architectures, particularly addressing the challenges of rare cell populations [20, 19]. Other methods have leveraged bulk expression data to identify phenotype-associated subpopulations [18]. While progress has been

© The Author 2022. Published by Oxford University Press. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com

made in patient classification using single-cell data, current approaches focus on interpretability at the cell type level rather than identifying condition-specific cell states within each type [22]. Moreover, these methods require extensive human intervention or additional data sources, limiting their practical application.

Multiple Instance Learning (MIL) offers a promising framework for addressing these challenges, having demonstrated success in various biological domains, including histopathological image analysis [4, 15, 23]. In the context of single-cell transcriptomics, MIL's bag-based learning paradigm naturally aligns with the hierarchical structure of single-cell data, where samples can be viewed as bags containing multiple cell instances. However, unlike other applications, single-cell transcriptomes present unique challenges due to the need for ground truth labels at the cell level.

To address these limitations, we propose scMILD (Single-cell Multiple Instance Learning for Sample Classification and Associated Subpopulation Discovery), a framework that leverages Multiple Instance Learning to identify condition-associated cell subpopulations while providing quantitative performance metrics through sample-level classification. By treating samples as bags and cells as instances, scMILD offers a systematic approach to bridge sample-level phenotypes with cell-level molecular signatures.

We validate scMILD using both simulation studies with CRISPR-perturbed cells and diverse disease datasets, including Lupus, COVID-19, and Ulcerative Colitis. Our results demonstrate scMILD's ability to identify biologically relevant cell subpopulations while maintaining robust sample classification performance. These findings suggest scMILD's potential as a reliable tool for exploring cellular heterogeneity in complex diseases, offering a more systematic and reproducible approach to condition-associated cell subpopulation discovery.

Materials

Simulation Dataset

To systematically evaluate model performance under controlled settings, we designed a simulation study using preprocessed single-cell RNA sequencing data from a CRISPRa-based gene over-expression experiment [9].

The preprocessed data, filtered to include only cells with good coverage as defined in [11], contains control and KLF1-perturbed cells, with perturbed cells serving as ground truth condition-specific cells to determine sample conditions. We identified 2000 Highly Variable Genes (HVGs) for feature selection using `scanpy` [21].

Disease Datasets

In this study, we evaluated the performance of scMILD using single-cell RNA-seq datasets representing various biological conditions. The datasets include samples from lupus patients, COVID-19 patient nasal swabs (Infection), peripheral blood mononuclear cells (Hosp.), and ulcerative colitis (UC) samples. Each dataset comprises a varying number of samples and cells, with detailed information summarized in Table 1.

We selected 3000 HVGs for feature selection using the same approach as the simulation dataset for the Lupus, COVID-19 Infection, and COVID-19 Hospitalization datasets. For the COVID-19 Infection dataset, an additional filtering step was applied to retain only genes expressed in at least five cells. The UC dataset utilized 2000 HVGs. Raw counts were used as model inputs across all datasets.

Table 1. Disease Dataset Information

Dataset[1]	Samples	Classes[2]	Cells	Avg. Cells[3]	Used Genes
Lupus [7]	169	SLE: 119 Healthy: 50	834,096	4,935	3,000
COVID-19 Infection [25]	50	Infection: 35 Normal: 15	26,947	539	3,000
COVID-19 Hosp. [17]	254	Hosp.: 144 Not-Hosp.: 110	515,141	2,028	3,000
UC [16]	42	Inflamed: 18 Healthy: 24	18,725	446	2,000

[1]Hosp.: Hospitalized; UC: Ulcerative Colitis.

[2]SLE: Systemic Lupus Erythematosus; Hosp.: Hospitalized; Not-Hosp.: Not Hospitalized.

[3]Avg. Cells: Average number of cells per sample.

Methods

scMILD

Encoder

The encoder aims to create a latent vector that effectively represents a cell's gene expression. The encoder function f_{enc} maps the gene expression vector $x_{i,j}$ to a latent representation $z_{i,j}$, i.e., $z_{i,j} = f_{\text{enc}}(x_{i,j})$, where $z_{i,j} \in \mathbb{R}^k$ and k is the dimensionality of the latent space. For cell i in sample j, the autoencoder's decoder outputs $\hat{x}_{i,j}$ and $\theta_{i,j}$ are used as parameters for a negative binomial distribution. The negative log-likelihood loss of the original gene expression counts $x_{i,j}$ for this distribution is calculated. The negative binomial probability mass function is:

$$\text{NB}(x|\hat{x},\theta) = \frac{\Gamma(x+\theta)}{\Gamma(\theta)}\left(\frac{\theta}{\theta+\hat{x}}\right)^{\theta}\left(\frac{\hat{x}}{\theta+\hat{x}}\right)^{x} \tag{1}$$

The overall loss for the autoencoder is calculated as:

$$\mathcal{L}_{NB} = \frac{1}{N}\sum_{j=1}^{M}\sum_{i=1}^{n_j}\sum_{g=1}^{G} -\log \text{NB}(x_{i,j,g}|\hat{x}_{i,j,g},\theta_{i,j,g}) \tag{2}$$

where $N = \sum_{j=1}^{M} n_j$ represents the total number of cells across all samples, with M denoting the number of samples, n_j the number of cells in sample j, and $g \in \{1,\ldots,G\}$ indexing genes with G being the total number of genes. The autoencoder is optimized by minimizing $\mathcal{L}_{NB}$ with respect to the encoder parameters ϕ and decoder parameters ψ:

$$\min_{\phi,\psi} \mathcal{L}_{NB} \tag{3}$$

After pretraining, the encoder f_{enc} with parameters ϕ is used as a shared branch for the sample and cell networks in the scMILD model.

Attention Module

The attention module function f_{att}, adapted from Attention based MIL [4], computes the attention score $a_{i,j}$ for each cell based on its latent representation $z_{i,j}$, i.e., $a_{i,j} = f_{\text{att}}(z_{i,j})$, where $a_{i,j} \in [0, 1]$. The attention scores are computed as follows:

$$a_{i,j} = \frac{\exp\left(\mathbf{w}^{\mathsf{T}} \tanh(\mathbf{V} z_{i,j}^{\mathsf{T}})\right)}{\sum_{l=1}^{n_j} \exp\left(\mathbf{w}^{\mathsf{T}} \tanh(\mathbf{V} z_{l,j}^{\mathsf{T}})\right)} \tag{4}$$

where $\mathbf{V} \in \mathbb{R}^{p \times k}$ and $\mathbf{w} \in \mathbb{R}^{p}$ are learnable parameters of the attention module, with p being a hyperparameter.

Sample Branch

The sample branch consists of the shared encoder f_{enc}, the attention module f_{att}, and the sample classifier $f_{\text{clf}}^{\text{sample}}$. The sample-level features v_j for sample s_j are obtained by taking a weighted sum of the cell latent representations using the cell attention scores:

$$v_j = \sum_{i=1}^{n_j} a_{i,j} z_{i,j} \tag{5}$$

The sample classifier $f_{\text{clf}}^{\text{sample}}$ takes the sample-level features v_j as input and outputs the predicted probability $\hat{y}_j$ of the sample belonging to the positive class, i.e., $\hat{y}_j = f_{\text{clf}}^{\text{sample}}(v_j)$. The binary cross-entropy loss for the sample branch is calculated as:

$$\mathcal{L}_{\text{sample}} = -\frac{1}{M} \sum_{j=1}^{M} \left(y_j \log(\hat{y}_j) + (1 - y_j) \log(1 - \hat{y}_j)\right) \tag{6}$$

Where y_j is the true label of sample s_j. During training, the encoder, attention module, and sample classifier of the sample branch are updated to minimize the binary cross-entropy loss:

$$\min_{\phi, \mathbf{V}, \mathbf{w}, \omega} \mathcal{L}_{\text{sample}} \tag{7}$$

Where ϕ, $\mathbf{V}$, $\mathbf{w}$, and ω represent the parameters of the encoder, attention module, and sample classifier, respectively.

Cell Branch

The cell branch comprises the shared encoder f_{enc} and the cell classifier $f_{\text{clf}}^{\text{cell}}$. The cell classifier takes the latent representation $z_{i,j}$ of a cell as input and outputs the predicted probability $\hat{\tau}_{i,j}$ of the cell belonging to the positive class, i.e., $\hat{\tau}_{i,j} = f_{\text{clf}}^{\text{cell}}(z_{i,j})$. The pseudo-label $\tau_{i,j}$ of the i-th cell in sample s_j is derived from the cell attention score $a_{i,j}$:

$$\tau_{i,j} = \begin{cases} a_{i,j}, & \text{if } y_j = 1 \\ 0, & \text{if } y_j = 0 \end{cases} \tag{8}$$

The weighted binary cross-entropy loss for the cell branch is calculated as:

$$\begin{aligned} \mathcal{L}_{\text{WCE}} = -\frac{1}{N} \sum_{j=1}^{M} \sum_{i=1}^{n_j} & \left(\lambda(1 - \tau_{i,j}) \log(1 - \hat{\tau}_{i,j})\right. \\ & \left. + (1 - \lambda)\tau_{i,j} \log(\hat{\tau}_{i,j})\right) \end{aligned} \tag{9}$$

where λ is the negative label weight, which controls the balance between the loss contributed by negative and positive pseudo-labels.

To further enhance the model's ability to distinguish between condition-associated and non-associated cells, we incorporate an orthogonal projection loss (OPL) [13]. This loss is applied to cells from positive samples ($y_j = 1$) and is calculated as follows:

1. We perform 2-component Gaussian mixture modeling on $\tau_{i,j}$ values of cells from positive samples to divide them into two clusters: C_{high} and C_{low}.

2. In the embedding space defined by the encoder, we calculate the normalized OPL as:

$$\begin{aligned} s &= \frac{1}{|P|} \sum_{c_i = c_j} \langle \mathbf{z}_i, \mathbf{z}_j \rangle \\ d &= \frac{1}{|N|} \sum_{c_i \neq c_k} \langle \mathbf{z}_i, \mathbf{z}_k \rangle \\ \mathcal{L}_{\text{OPL}} &= (1 - s) + |d| \end{aligned} \tag{10}$$

Here, $|P|$ and $|N|$ denote the numbers of positive and negative pairs respectively, and c_i represents the cluster label (high or low) assigned by GMM for cell i. The cosine similarity operator $\langle \cdot, \cdot \rangle$ between two cell embedding vectors is defined as:

$$\langle \mathbf{z}_i, \mathbf{z}_j \rangle = \frac{\mathbf{z}_i \cdot \mathbf{z}_j}{\|\mathbf{z}_i\|_2 \cdot \|\mathbf{z}_j\|_2} \tag{11}$$

where $\|\cdot\|_2$ refers to the ℓ_2 norm operator. This formulation encourages the model to learn similar embeddings for cells within the same cluster while maintaining orthogonality between cells from different clusters, thereby enhancing the separation between condition-associated and non-associated cells within positive samples.

The final loss function for the cell branch is a combination of the weighted binary cross-entropy loss and the OPL:

$$\mathcal{L}_{\text{cell}} = \mathcal{L}_{\text{WCE}} + \gamma \mathcal{L}_{\text{OPL}} \tag{12}$$

Where γ is a hyperparameter controlling the contribution of the OPL.

During training, the encoder and cell classifier of the cell branch are updated to minimize the total loss:

$$\min_{\phi, \psi} \mathcal{L}_{\text{cell}} \tag{13}$$

Where ϕ and ψ represent the parameters of the encoder and cell classifier, respectively.

Incorporating the OPL into the cell branch's loss function enhances the model's ability to identify condition-associated cell subpopulations. This approach encourages a more apparent separation between high-attention and low-attention cells in the embedding space for positive samples. This leads to more effective learning of sample condition-associated cellular patterns.

Training Procedure

The training procedure consists of three main stages: pretraining the autoencoder, optimizing the sample branch, and optimizing the cell branch.

Pretraining the Autoencoder

The autoencoder is pretrained to create a latent representation of cell gene expression, minimizing the negative binomial loss (2). Key hyperparameters include encoder layers (512, 256, 128), learning rate (0.001), max epochs (250), and early stopping patience (15). Batch size is set to 128 for all datasets.

Optimizing the Sample-Cell Branch

scMILD optimization alternates between Sample and Cell Branch:

1. Sample Branch:
   - Compute sample-level features and predict sample labels.
   - Update parameters using binary cross-entropy loss (6).
2. Cell Branch:
   - Predict cell labels and derive pseudo-labels from attention scores.
   - Apply Gaussian mixture modeling to positive samples.
   - Calculate and combine weighted binary cross-entropy loss (9) and orthogonal projection loss (10).
   - Update parameters using the total cell branch loss (12).

The sample-cell optimization ratio varies across datasets (1 for most, 5 for COVID-19 Hospitalization). Other hyperparameters include the Adam optimizer, learning rates, dimensions, and early stopping criteria. For detailed hyperparameter settings, see Supplementary Table 2.

Experimental Setup

We adopted a rigorous evaluation framework for experimental validation following ProtoCell4P [22]. Each dataset was systematically partitioned into training (50%), validation (25%), and test (25%) sets. We conducted eight independent runs with different random seeds to ensure statistical robustness and implemented early stopping based on validation metrics.

We performed a comprehensive comparative analysis against state-of-the-art models WENO [12] and ABMIL [4]. All models utilized identical preprocessed data and the same pretrained autoencoder architecture to ensure a fair comparison. Hyperparameters, including learning rates, hidden dimensions, and optimization strategies, were maintained consistently across models where applicable (detailed settings in Supplementary Table 2).

Common Analysis Methods

Cell Stratification and Condition-association Grouping

For systematic cell population analysis, we employed the attention scores from scMILD's attention module to perform cell stratification. Within each test dataset, we applied a 2-component Gaussian mixture model (GMM) to partition cells based on their attention scores, mirroring the approach used in the Cell Branch during model training.

Based on this GMM clustering, we defined two groups:

1. Condition-associated group: Cells that belong to the cluster with higher mean attention scores and are derived from Positive samples (e.g., Infected, Inflamed, or Hospitalized).

2. Condition-independent group: Cells belonging to the cluster with lower mean attention scores derived from Positive samples.

This stratification strategy enabled focused analysis of cellular heterogeneity within Positive samples, allowing discrimination between condition-associated and condition-independent cellular states. We intentionally excluded cells from Negative samples from this categorization to maintain an analytical focus on understanding disease-relevant cellular heterogeneity.

Differential Gene Expression Analysis

To identify differentially expressed genes (DEGs) between cell groups, the `FindMarkers` function from `Seurat` was used [3]. A log2 fold change >0.25 and an adjusted p-value <0.01 were set as cutoffs.

Simulation Study Design

We developed a systematic evaluation framework using the simulation dataset to assess scMILD's performance under controlled settings. We constructed binary classification datasets comprising two groups of 14 samples each: control samples containing only control cells and case samples containing both control and perturbed cells at varying ratios. This design enabled rigorous evaluation of sample classification accuracy and cell-level identification capability.

Dataset Construction

To ensure realistic evaluation parameters, we established a baseline of 400 cells per sample, derived from the minimum average cell count observed in our disease datasets (UC: 446 cells/sample). We then systematically evaluated model performance with decreasing cell counts (300, 200, 100) to assess robustness under data-limited scenarios.

The proportion of perturbed cells in case samples was carefully calibrated based on previously documented disease-associated cell subpopulations in SLE patients [8]. The baseline proportion was set to 20%, reflecting the observed average (21.2%) across various cell types: dendritic cells (cDC-caSC2: 25.5%, pDC-caSC3: 15.9%), NK cells (NK-caSC3: 31.2%), monocytes (Mono-caSC0: 35.4%, Mono-caSC4: 9.6%), and T cells (T-caSC6: 9.5%). To evaluate detection sensitivity across different abundance levels, we included additional proportions of 10% and 5%. Table 2 provides a comprehensive overview of the simulation parameters.

Table 2. Simulation Dataset Parameters

Parameter	Values
Number of samples per group	14
Cells per sample	400, 300, 200, 100
Proportion of perturbed cells	20%, 10%, 5%
Total experimental settingss	12 (4 cell counts × 3 proportions)

Note: Control samples contain only control cells, while case samples contain both control and KLF1-perturbed cells at specified proportions.

Evaluation Strategy

Our evaluation framework incorporated both sample-level and cell-level performance metrics. Sample-level assessment utilized AUROC scores across all experimental configurations, while cell-level evaluation employed min-max scaled attention scores for perturbed cell identification, validated against ground truth labels.

To establish the robustness of scMILD's cell identification capabilities, we implemented a comprehensive distributional analysis of attention scores between models. Each experimental configuration was evaluated using Kolmogorov-Smirnov (KS) statistics and overlap coefficients, comparing attention score distributions between perturbed and control cells. These metrics were aggregated across eight independent experimental runs to ensure statistical reliability.

For the median-performing experiment from each setting, we evaluated the ability to detect true differential expression signals through three complementary approaches: 1) Ground truth comparison: analyzing differentially expressed genes between KLF1-perturbed vs control cells 2) scMILD-based analysis: comparing case-associated vs other cells based on GMM clustering of attention scores 3) Phenotype-based analysis: comparing cells from case vs control samples

Disease Dataset Analysis

Performance Evaluation

We evaluated scMILD's performance on disease datasets using two complementary approaches. We calculated AUROC and F1-macro scores averaged across eight independent runs for sample-level classification. We applied the same stratification and differential expression methods validated in our simulation study for cell-level analysis, with adaptations specific to each disease context.

Calculation of Known Subcluster/Subtype Signature Scores

We calculated signature scores for known cell subclusters and subtypes associated with each condition to evaluate the biological relevance of the identified cell subpopulations.

For the Lupus dataset, we utilized the top 100 differentially expressed genes (DEGs) for each known ISG^{hi} SLE-expanded subcluster, as provided in the original study [8]. Signature scores were computed using the `MetaFeature` function from the `Seurat` package [3], which calculates the average expression of each gene set subtracted by the aggregated expression of control feature sets.

For the COVID-19 infection dataset, we employed the `UCell` package [1] to calculate signature scores for known cell subtypes. Specifically, we used the `AddModuleScore_UCell` function to compute scores for each cell subtype using their top 5 marker genes, as identified in the original study [25]. `UCell` implements a rank-based scoring method robust to batch effects and differences in sequencing depth.

These signature scores allowed us to quantitatively assess the enrichment of known condition-associated gene signatures in our identified cell subpopulations, providing a means to validate and interpret the biological significance of scMILD's results.

Pseudotime Analysis

Pseudotime analysis was performed on the UC dataset using the `learn_graph` function from `Monocle3` [2]. Default parameters were used. The principal node of the inferred graph with low cell attention scores or a high distribution of healthy cells was selected as the root node. The Spearman correlation between min-max scaled cell attention scores and pseudotime was calculated.

Results

Model Architecture and Training Strategy

The scMILD architecture comprises two main components: a Sample Branch and a Cell Branch (Figure 1a). The model operates on two key assumptions: (1) a positive condition sample contains at least one positive cell, and (2) most cells in a positive sample are negative. Based on these assumptions, scMILD uses attention scores from the Sample Branch as pseudo-labels for cells in positive samples while treating all cells from negative samples as true negatives for Cell Branch training.

To establish the model foundation, the Sample Branch performs initial learning through sample-level classification. It employs an attention mechanism to identify potentially condition-associated cells based on their contribution to sample classification. These attention scores then serve as pseudo-labels for the Cell Branch, which refines the model through cell-level prediction (Figure 1a). The interaction between branches is optimized through an alternating training process where each branch leverages the other's progress (Figure 1b). This process is enhanced by integrating a Gaussian mixture model and orthogonal projection loss in the Cell Branch, improving the separation between condition-associated and non-associated cells.

The model optimization involves sequential updates of the Sample and Cell branches. While the Sample Branch focuses on accurate sample classification, the Cell Branch refines cellular state representations using the pseudo-labels derived from attention scores. This dual optimization strategy ensures both robust sample-level predictions and meaningful cell-level representations.

This integrated approach allows scMILD to effectively leverage cellular heterogeneity for robust sample classification and disease-associated cell subpopulation discovery in single-cell transcriptomic data.

Simulation Study Results

To objectively evaluate scMILD's performance and practical utility in biological research, we designed a controlled simulation study where ground truth cell labels were available through CRISPR-mediated perturbation (Figure 2a). This experimental setup allowed us to assess both sample- and cell-level classification performance while evaluating the model's capability to identify biologically meaningful signals under various conditions.

Model Performance Analysis

Analysis of sample classification performance revealed scMILD's superior robustness across different experimental settings (Figure 2b). With 400 cells and 20% perturbation, scMILD achieved perfect sample classification (AUROC = 1.0), outperforming both WENO (AUROC = 0.989) and ABMIL (AUROC = 0.833). This performance advantage was maintained even under more challenging conditions, with scMILD consistently demonstrating higher AUROC values across different cell counts and perturbation ratios.

The cell-level performance showed similar trends, with scMILD demonstrating a superior ability to identify perturbed cells. At 400 cells and 20% perturbation, scMILD achieved a cell-level AUROC of 0.956, compared to 0.905 for WENO and 0.615 for ABMIL. Notably, scMILD maintained strong performance despite reduced cell counts and perturbation rates.

The relationship between sample and cell-level classification revealed a stronger correlation in scMILD (r = 0.892) compared to WENO (r = 0.879) and ABMIL (r = 0.556) (Figure 2c). This stronger correlation indicates that sample-level classification performance could serve as a reliable indicator of cell-level classification accuracy in real-world applications where ground truth cell labels are unavailable, providing a practical way to assess the model's cell-level performance through sample-level validation.

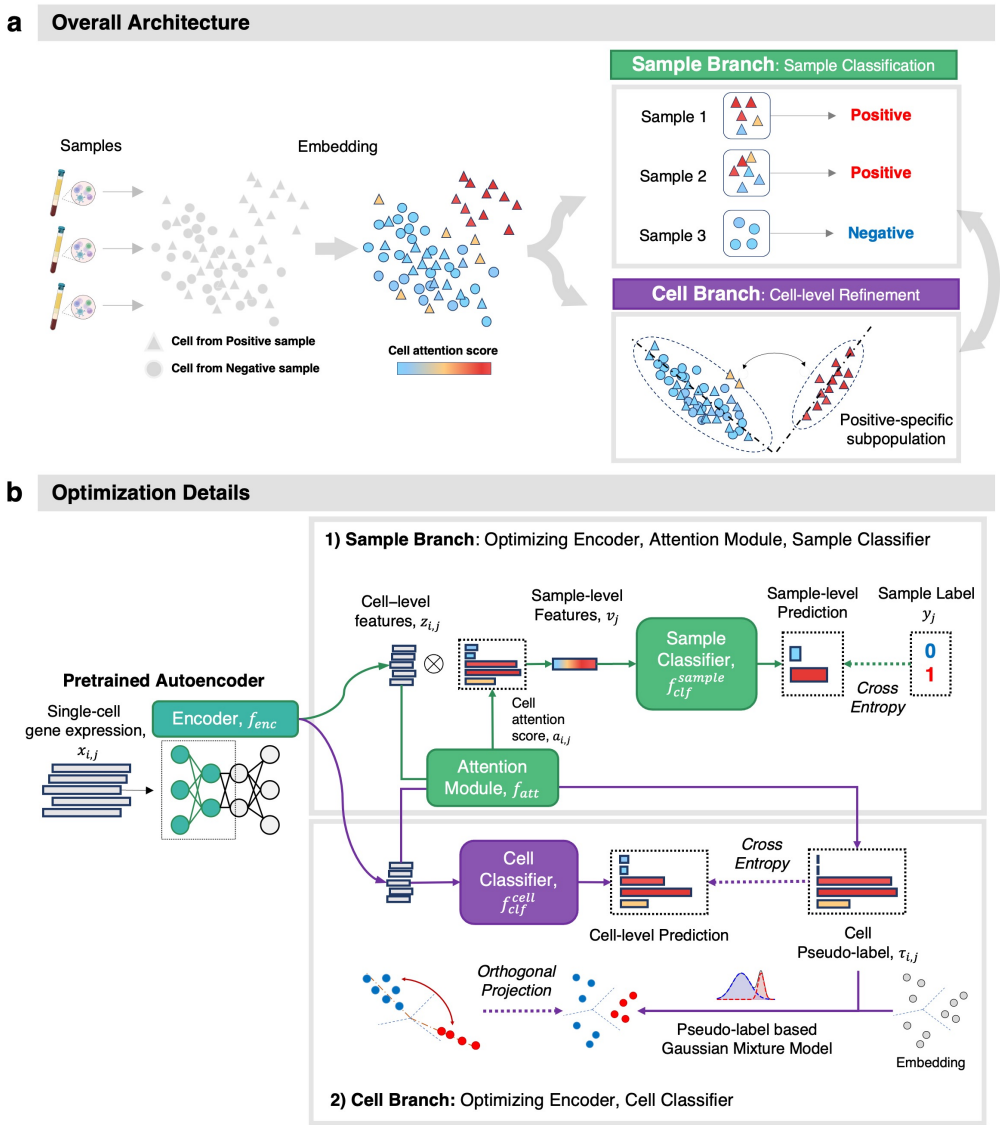

Fig. 1. Overview of the scMILD framework. **a**, The scMILD model architecture and training process. The Sample Branch performs initial learning using sample-level classification to establish the model foundation, while the Cell Branch refines the model using cell-level prediction with pseudo-labels derived from attention scores. **b**, Optimization details of the Sample and Cell Branches.

Cell Attention Score Analysis

Analysis of the attention score distributions between KLF1-perturbed and control cells revealed higher Kolmogorov-Smirnov statistics and lower overlap coefficients for scMILD compared to other models (Supplementary Table 3). This improved separation was evident in the split violin plots (Figure 2d), where scMILD demonstrated a more apparent distinction between perturbed and control cell distributions across different cell counts (Supplementary Figure 1). Furthermore, UMAP visualization of the median-performing experiment (400 cells, 20% perturbation) showed scMILD's ability to capture cellular state differences (Figure 2e-f), with cell attention scores showing better alignment with the ground truth perturbation pattern compared to other models. Additional UMAP visualizations for experiments with lower cell counts are provided in Supplementary Figure 2.

Differential Expression Analysis

Using cell attention scores, we applied Gaussian mixture modeling to categorize cells into condition-associated and non-associated groups. To evaluate the practical utility of these identified groups, we compared scMILD's ability to detect differentially expressed genes with phenotype-based analysis, using ground truth from KLF1-perturbed cells as a reference.

While both approaches showed high precision (scMILD: 0.87-0.91, Phenotype-based: 0.93-1.0), scMILD identified substantially more DEGs with higher recall rates (scMILD: 0.51-0.92, Phenotype-based: 0.0-0.14). At 400 cells per sample, scMILD detected 150 DEGs with 130 true positives, compared to only 14 DEGs in the phenotype-based analysis. Notably, when cell counts were low (100 cells), phenotype-based analysis failed to detect any DEGs, while scMILD maintained reliable detection capability, identifying 39 DEGs with 34 true positives.

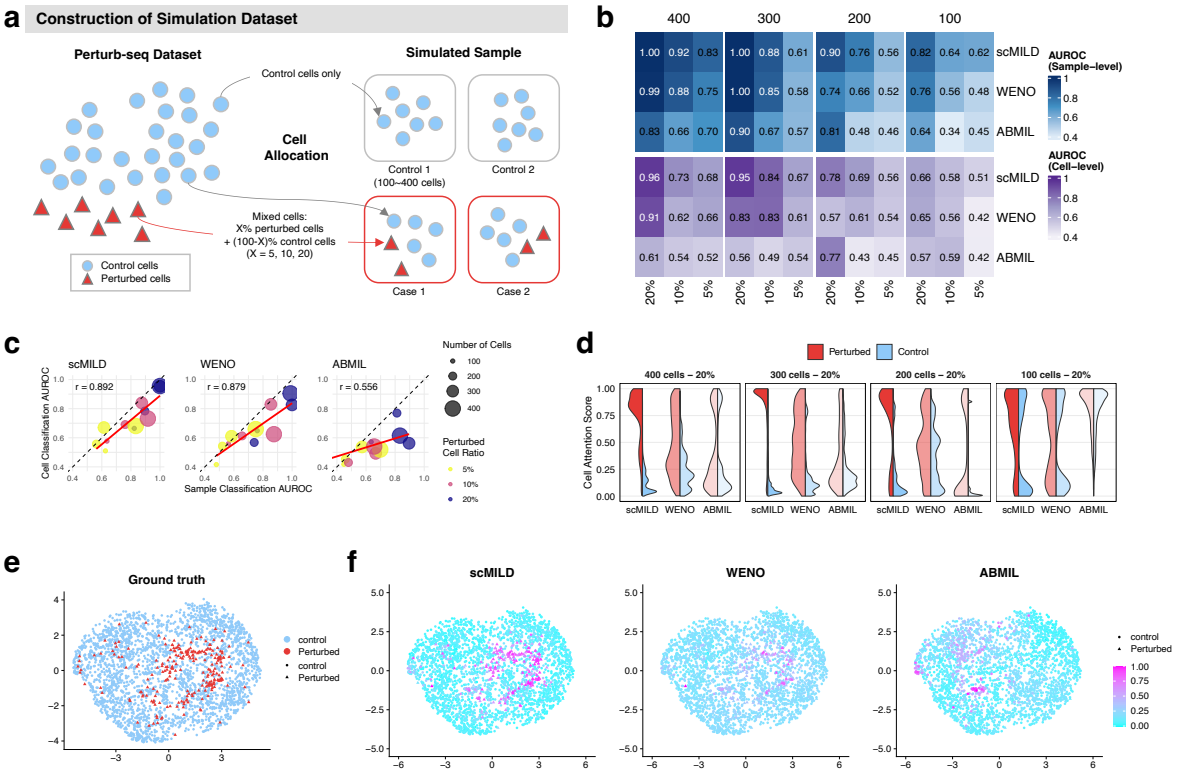

Fig. 2. **a**, Schematic of simulated dataset construction from Perturb-seq data, with control (blue circles) and perturbed cells (red triangles) allocated to control and case samples. **c**, Scatter plot comparing sample and cell classification AUROC, with point size reflecting total cell number and color indicating perturbed cell fraction. **d**, Split violin plots of cell attention score distributions from the three models for perturbed (red) and control (blue) cells, with 20% perturbed cells and varying total cell numbers. **e-f**, UMAP visualizations of the median-performing test set showing ground truth cell labels (**e**) and cell attention scores from each model (**f**), with perturbed cells represented as triangles and control cells as circles.

These results demonstrate scMILD's ability to effectively capture biological signals in datasets with limited cell numbers, providing a robust approach for identifying condition-associated molecular signatures.

Table 3. Comparison of DEG identification performance

Cells/ Sample	Ground Truth DEGs	scMILD		Phenotype-based	
		DEGs (TP[1]/Total)	Prec/Rec[2]	DEGs (TP/Total)	Prec/Rec
400	157	130/150	.87/.83	14/14	1.0/.09
300	103	75/83	.90/.73	14/15	.93/.14
200	63	32/35	.91/.51	5/5	1.0/.08
100	37	34/39	.87/.92	0/0	-/0.0

[1]TP: True Positives (overlap with ground truth DEGs).

[2]Prec/Rec: Precision/Recall.

Performance Evaluation on Disease Datasets

To evaluate scMILD's effectiveness, we compared its performance with state-of-the-art models WENO[12], ABMIL[4], and ProtoCell4P[22] on four single-cell RNA-seq datasets: Lupus, COVID-19 Infection, COVID-19 Hospitalization (Hosp.), and Ulcerative Colitis (UC). For a fair comparison, we used an identical pre-trained autoencoder for scMILD, WENO, and ABMIL, ensuring that differences in performance were due to the models' unique architectures.

We assessed performance using AUROC and F1-macro scores. As shown in Table 4, scMILD demonstrated superior or comparable performance across all datasets and metrics.

In the Lupus dataset, scMILD achieved perfect AUROC (1.0000) alongside ProtoCell4P while outperforming other models in the F1-macro score (0.9801). For COVID-19 Infection, scMILD surpassed all models in both AUROC (0.9063) and F1-macro (0.7769) scores. Similarly, in COVID-19 Hospitalization, scMILD achieved the highest scores (AUROC: 0.9562, F1-macro: 0.9020).

The most significant performance difference was observed in the UC dataset, where scMILD substantially outperformed all models (AUROC: 0.9750, F1-macro: 0.9001). This is particularly noteworthy given the UC dataset's inflammation-specific cell types, suggesting scMILD's ability to capture disease-specific cellular patterns without explicit cell type information.

Notably, scMILD consistently outperformed WENO and ABMIL across all datasets, indicating that its structural innovations, particularly the orthogonal projection loss and the cell branch, contribute significantly to improved performance.

Table 4. Performance comparison of scMILD with WENO, ABMIL, and ProtoCell4P on various single-cell RNA-seq datasets.

	AUROC				F1-macro			
Dataset	scMILD	WENO[12]	ABMIL[4]	ProtoCell4P[22]	scMILD	WENO[12]	ABMIL[4]	ProtoCell4P[22]
Lupus	**1.0000**	0.9997	0.9968	**1.0000**	**0.9801**	0.9765	0.9635	0.9742
COVID-19 Infection	**0.9063**	0.8854	0.8646	0.8493	**0.7769**	0.7562	0.6620	0.7097
COVID-19 Hosp.	**0.9562**	0.9529	0.9555	0.9549	**0.9020**	0.8880	0.8957	0.8754
UC	**0.9750**	0.9583	0.8917	0.9370	**0.9001**	0.8509	0.8514	0.5941

Note: The best performance for each dataset and metric is highlighted in bold.

These results demonstrate scMILD's strength in single-cell transcriptome-based sample classification tasks and its potential for accurately classifying samples and identifying disease-associated cell subpopulations in complex diseases. Detailed performance comparisons are provided in Supplementary Table 1.

Validation of scMILD-identified Condition-associated Cell Subpopulations

To validate the effectiveness of scMILD in identifying condition-associated cell subpopulations, we applied our model to four distinct datasets representing different disease conditions: Lupus, COVID-19 infection, COVID-19 hospitalization, and Ulcerative Colitis. For each dataset, we conducted downstream analysis on the median-performed test set. Our approach involved applying a 2-component Gaussian mixture model based on cell attention scores to divide cells derived from Positive condition samples into two subgroups. The subgroup with higher scores was designated as the condition-associated subgroup, while the other was termed the condition-independent subgroup.

We then focused on identifying and characterizing the condition-associated subgroup, comparing it with known subtypes or markers reported in previous studies. This systematic approach allowed us to evaluate scMILD's performance across diverse pathological contexts and validate our findings against existing knowledge. Our analysis demonstrates the model's capability to detect biologically relevant cell subtypes across various disease conditions, as detailed in the following sections.

Lupus: Identification of SLE-associated Cell Subtypes

The Lupus dataset comprised 50 samples from healthy individuals and 119 samples from systemic lupus erythematosus (SLE) patients. Figure 3a and b present UMAP visualizations of the dataset, providing insights into cellular heterogeneity and SLE-associated changes. In Figure 3a, we observe distinct clusters corresponding to different cell types, with notable populations of B, T, and monocytes. Figure 3b reveals the distribution of Healthy, SLE-independent, and SLE-associated subgroups within these cell type clusters. Interestingly, the SLE-associated cells are not confined to specific regions but are distributed across multiple cell types, suggesting that SLE-associated changes occur within various cell populations.

Our analysis revealed that CD14+ Monocytes and FCGR3A+ Monocytes exhibited the highest cell attention scores, consistent with previous research findings on SLE-associated cell types (Figure 3c). This observation aligns with the known involvement of monocytes in SLE pathogenesis.

To further investigate cellular heterogeneity, we employed Gaussian mixture modeling on cell attention scores, independent of cell type, to divide cells from SLE patients into "SLE-associated" and "SLE-independent" subgroups. We then compared our findings with previously reported SLE-associated cell subtypes.

We created a marker gene expression score by leveraging the 100 marker genes for each ISG^{hi} SLE-expanded subcluster reported in a previous study [8]. Notably, the ISG15 gene, common to all subclusters, showed a high Pearson correlation coefficient of 0.5997 with our cell attention scores (Figure 3d). The feature plot in Figure 3e further illustrates the expression pattern of ISG15 across cells in the UMAP space, clearly showing higher expression in regions corresponding to the SLE-associated subgroup.

Furthermore, we compared the expression of meta-features (Top 100 marker genes of ISG^{hi} SLE-expanded subcluster) across Healthy, SLE-independent, and SLE-associated subgroups for different cell types (Figure 3f). The top 100 differentially expressed genes for each ISG^{hi} SLE-expanded subcluster used in this analysis are listed in Supplementary Table 4. The box plots reveal a consistent pattern across all examined cell types: the SLE-associated subgroup demonstrates significantly higher expression of these marker genes than the Healthy and SLE-independent subgroups. This trend is observed across various cell types, including monocytes, dendritic cells, and lymphocytes.

These results validate scMILD's ability to identify disease-relevant cell subpopulations in SLE, aligning with and extending previous findings in the field. The consistency of our results across different cell types and the clear distinction in meta-feature expression between subgroups underscore the robustness of our approach in capturing disease-associated cellular states.

COVID-19 Infection: Detection of Virus-specific Cellular Responses

The COVID-19 infection dataset, which includes measurements of SARS-CoV2 mRNA expression, provided insights into cellular heterogeneity and infection-associated changes. Figure 4a and b present UMAP visualizations of the dataset, revealing distinct cell type clusters and the distribution of Normal, Infection-independent, and Infection-associated subgroups within these clusters.

The Infection-associated cells were distributed across multiple cell types rather than forming a separate cluster. These cells tended to be located towards the periphery of their respective cell type clusters, suggesting that they represent states deviating from typical cellular profiles in normal conditions. Figure 4c quantitatively demonstrates the enrichment of SARS-CoV2 mRNA-detected cells in the infection-associated subgroup across all cell types, supporting the biological relevance of our identified subgroup.

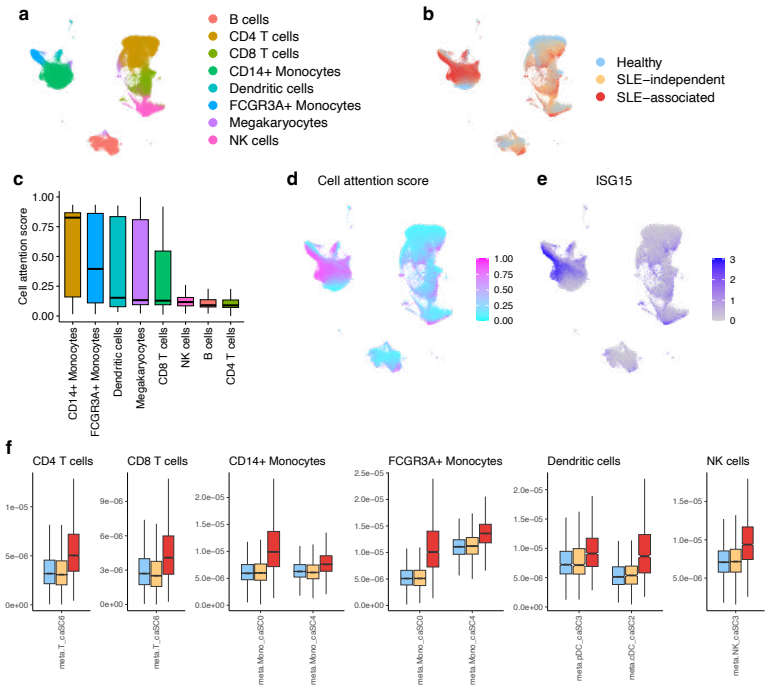

Fig. 3. **a**, UMAP visualization of the Lupus dataset, colored by cell type. **b**, UMAP visualization of the Lupus dataset, colored by SLE association subgroup (Healthy, SLE-independent, and SLE-associated). **c**, Box plot showing the distribution of cell attention scores for each cell type. **d**, Feature plot of cell attention scores. **e**, Feature plot showing the expression of ISG15 across cells in the UMAP space. **f**, Box plots comparing meta-feature (ISG^{hi} SLE-expanded subcluster marker gene) expression scores across Healthy, SLE-independent, and SLE-associated subgroups for different cell types.

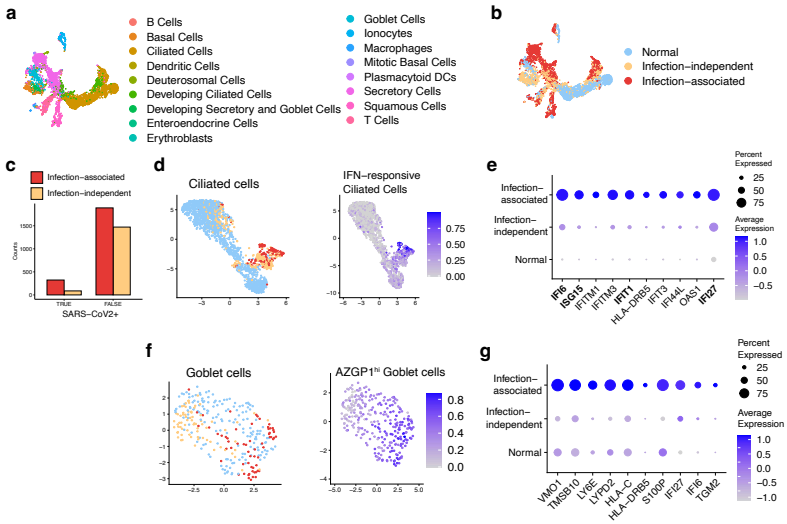

Fig. 4. **a**, UMAP visualization of the COVID-19 infection dataset, colored by cell type. **b**, UMAP visualization of the dataset, colored by infection association subgroup (Normal, Infection-independent, and Infection-associated). **c**, Bar plot showing the proportion of SARS-CoV2 mRNA-detected cells in each subgroup. **d**, UMAP plot of infection association subgroup and Interferon-responsive Ciliated Cell subtype signature scores for Ciliated Cells. **e**, Dot plot of top 10 differentially expressed genes in the Infection-associated subgroup of Ciliated Cells. **f**, UMAP plot of infection association subgroup and $AZGP1^{hi}$ Goblet cell subtype signature scores. **g**, Dot plot of top 10 differentially expressed genes in the Infection-associated subgroup of Goblet Cells.

To validate our findings, we compared them with COVID-19 expanded cell subtypes reported by Ziegler et al. [25]. We focused on Ciliated Cells (1571:541, Normal:Infection) and Goblet cells (162:134, Normal:Infection), which had more than 100 cells in both normal and infection samples (Supplementary Table 5).

For Ciliated Cells, we examined the distribution of the Interferon-responsive Ciliated Cell subtype, previously reported as expanded in COVID-19 patients. Using the top 5 DEG

markers (IFI6, ISG15, IFIT1, IFI27, MX1) for this known subtype, we observed a strong correlation between the subtype signature score and our infection-associated subgroup (Figure 4d). The Pearson correlation coefficient between the known subtype signature score and our cell attention score was remarkably high at 0.6348. Moreover, all five known subtype markers were present in the DEGs of our infection-associated subgroup, with four of them (excluding MX1) appearing in the top 10 DEGs (Figure 4e). The dot plot in Figure 4e clearly shows the upregulation of these interferon-responsive genes in the Infection-associated subgroup compared to the Normal and Infection-independent subgroups.

For Goblet cells, we analyzed the AZGP1^{hi} goblet cell subtype, another COVID-19 expanded cell subtype reported in the original study. Due to the very low p-values of the DEGs provided in the original paper, we selected the top 30 genes to calculate the signature score for this cell subtype. Similar to our findings in Ciliated cells, we observed a strong association between this signature score and our infection-associated subgroup (Figure 4f). The Pearson correlation coefficient between the signature score and our cell attention score was high at 0.6564. Figure 4g shows the top 10 differentially expressed genes in the Infection-associated subgroup of Goblet Cells, highlighting these cells' distinct gene expression profiles in response to infection. Notably, many of these genes, such as LY6E, IFI6, and IFITM3, are known to be involved in the interferon response, further confirming the biological relevance of our identified subgroups.

These results demonstrate scMILD's ability to effectively identify virus-specific cellular responses in COVID-19 infection, corroborating previous studies while providing additional granularity in cell subtype identification.

COVID-19 Hospitalization: Characterization of Severity-associated Cell Populations

We classified cells into Hosp-associated, Hosp-independent, and Mild subgroups in the COVID-19 Hospitalization PBMC dataset, comprising 254 COVID-19 PBMC samples. Figure 5a and b present UMAP visualizations of the dataset, revealing distinct cell type clusters and precise separation between the hospitalization association subgroups. The UMAP plots demonstrate that certain cell types, mainly monocytes, show a higher proportion of Hosp-associated cells, suggesting their potential role in disease severity.

Most cells in the Hosp-associated subgroup were identified as Monocytes at the broad cell type level, with CD14+ monocytes being the predominant cell type at a more specific level. This is clearly illustrated in Figure 5c, where the stacked bar plot shows a higher proportion of Hosp-associated cells in the Mono (Monocyte) category than other cell types.

Our analysis revealed a striking similarity between the Hosp-associated subgroup and the "Dysfunctional CD14 Monocytes" subtype reported in the original study [17]. All marker genes for this subtype (S100A8, S100A12, S100A9, S100A6, and TSPO) were among the top 15 differentially expressed genes in our Hosp-associated subgroup within the CD14+ monocytes cell type. Figure 5d presents a dot plot of these differentially expressed genes, where the size of the dots represents the percentage of cells expressing the gene and the color intensity indicates the average expression level. The plot clearly shows the upregulation of these marker genes in the Hosp-associated subgroup compared to the Hosp-independent and Mild subgroups, particularly for genes like S100A8, TSPO, S100A9, and S100A12.

These results validate scMILD's ability to identify severity-associated cell populations in COVID-19, aligning closely with previous findings and potentially offering new insights into the cellular basis of disease severity. Identifying these dysfunctional monocytes as a critical feature of severe COVID-19 cases highlights the potential of scMILD in uncovering clinically relevant cellular subpopulations in complex diseases.

Ulcerative Colitis: Identification of Inflammation-specific Fibroblasts

In the Ulcerative colitis dataset, we focused on inflammatory fibroblasts, a cell type-specific to the inflamed condition (Figure 6a,b). Our analysis confirmed that scMILD appropriately assigned high cell attention scores to this cell type, validating the model's ability to recognize disease-specific cell populations (Figure 6c). Interestingly, the WNT2B+ Fos-lo 2 cell type exhibited the second-highest median cell attention score, with a wide distribution. Further investigation through pseudotime analysis revealed a strong correlation between pseudotime and cell attention scores within this cell type, with a Pearson correlation coefficient of 0.6 (Figure 6d). This correlation suggests that the cell attention scores assigned by scMILD may reflect these cells' developmental or activation trajectory in the context of ulcerative colitis.

These findings corroborate the original study's identification of inflammatory fibroblasts as a hallmark of inflamed tissue in ulcerative colitis and provide new insights into the potential developmental trajectories of disease-associated cell types.

In summary, across these four diverse disease contexts, scMILD consistently demonstrated its ability to identify condition-associated cell subpopulations that align with and extend previous findings. Our analysis revealed several key strengths of the scMILD approach: 1. Robustness: scMILD effectively identified relevant cell subpopulations across various pathological conditions, from autoimmune diseases to viral infections and inflammatory disorders. 2. Biological relevance: The condition-associated subgroups identified by scMILD showed strong correlations with known disease-specific markers and subtypes reported in previous studies. 3. Novel insights: Besides corroborating existing knowledge, scMILD provided new perspectives on cellular heterogeneity in disease contexts, such as the potential developmental trajectories of disease-associated cell types in ulcerative colitis. 4. Versatility: scMILD's performance across different tissue types (e.g., PBMCs, nasal epithelium, colon) demonstrates its adaptability to various biological systems. 5. Granularity: The model's ability to identify condition-associated cells within specific cell types offers a more nuanced understanding of disease processes at the cellular level.

These results demonstrate scMILD's robustness and potential for discovering novel insights into cellular heterogeneity in various pathological conditions. By leveraging cell attention scores to identify associated subpopulations, scMILD provides a powerful tool for uncovering the cellular basis of complex diseases, potentially leading to new avenues for therapeutic interventions and personalized medicine approaches.

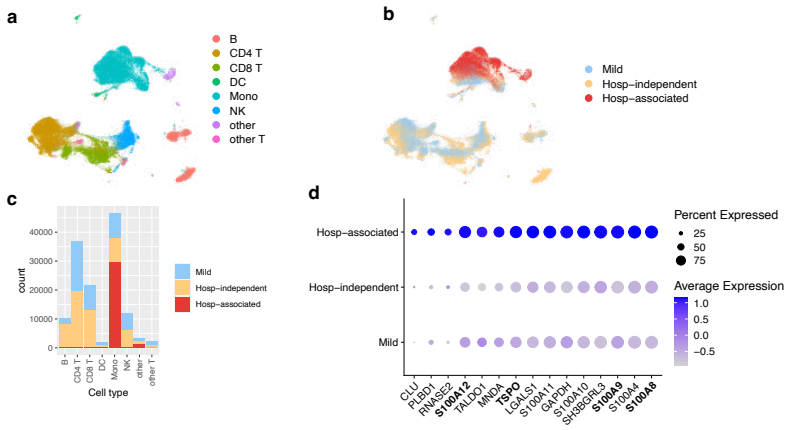

Fig. 5. **a**, UMAP visualization of the COVID-19 Hospitalization dataset, colored by cell type. **b**, UMAP visualization of the dataset, colored by hospitalization association subgroup (Mild, Hosp-independent, and Hosp-associated). **c**, Stacked bar plot displaying the number of cells in each subgroup for different cell types. **d**, Dot plot showing the expression levels and percentage of cells expressing the top differentially expressed genes in the Hosp-associated CD14+ Monocytes subgroup compared to other subgroups.

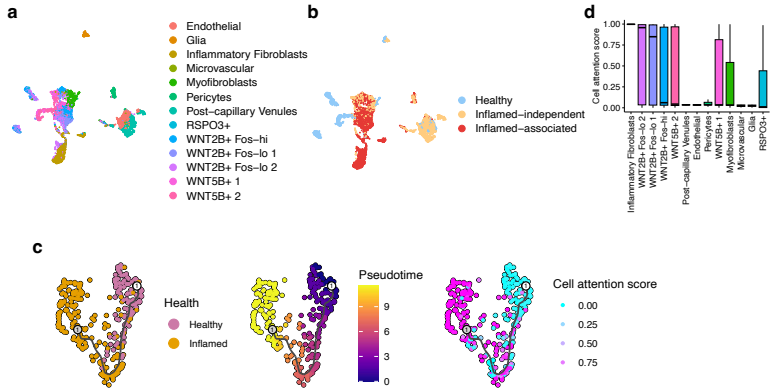

Fig. 6. **a-b**, UMAP visualization of the UC test dataset colored by cell type and subgroup. **c**, Distribution of cell attention scores for each cell type according to sample condition. **d**, Trajectory analysis results for WNT2B+ Fos-lo 2. Each column is colored by sample condition, pseudotime, and cell attention score.

Discussion

The model's effectiveness stems from its dual-branch architecture, where the Sample Branch performs robust sample-level classification while the Cell Branch identifies condition-associated cell subpopulations through enhanced representation learning. This integrated approach enables direct linking between cellular states and sample-level phenotypes.

We further validated scMILD's cell identification capabilities across varying cell counts and perturbation ratios using controlled simulation studies with CRISPR-perturbed cells. The model maintained reliable differential gene expression detection even with limited cell numbers where conventional approaches failed, demonstrating its technical robustness.

The biological relevance of scMILD was confirmed through its application to disease datasets, where our results aligned with findings from original studies. This consistency, along with new insights revealed by the model, positions scMILD as a crucial tool for investigating cellular heterogeneity in pathological conditions.

Future developments could extend scMILD to handle multi-class classification problems, enhancing its applicability to more complex biological scenarios. Improving result interpretability through explainable AI methodologies and incorporating pre-existing cell type annotations could provide deeper biological insights and reveal cell type-specific responses to pathological conditions.

In conclusion, scMILD represents a significant advancement in single-cell transcriptomic data analysis. By leveraging Multiple Instance Learning and its dual-branch architecture, scMILD offers a practical framework for uncovering cellular heterogeneity in complex biological systems. As we refine this approach, scMILD has the potential to enhance our understanding of cellular heterogeneity in health and disease, potentially leading to new therapeutic interventions and personalized medicine approaches.

Competing interests

No competing interest is declared.

Author contributions statement

K.J. contributed to data preprocessing, design and implementation of deep learning models, model training and optimization, and performance evaluation. K.K. (Co-corresponding author) contributed to the conceptualization of the study, methodology design, and supervision of model development. J.C. (Co-corresponding author) contributed to the overall research design, supervision of data analysis, and manuscript writing and revision.

Acknowledgments

The authors thank the anonymous reviewers for their valuable suggestions.

References

1. Massimo Andreatta and Santiago J Carmona. Ucell: Robust and scalable single-cell gene signature scoring. *Computational and structural biotechnology journal*, 19:3796–3798, 2021.
2. Junyue Cao, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M Ibrahim, Andrew J Hill, Fan Zhang, Stefan Mundlos, Lena Christiansen, Frank J Steemers, et al. The single-cell transcriptional landscape of mammalian organogenesis. *Nature*, 566(7745):496–502, 2019.
3. Yuhan Hao, Tim Stuart, Madeline H Kowalski, Saket Choudhary, Paul Hoffman, Austin Hartman, Avi Srivastava, Gesmira Molla, Shaista Madad, Carlos Fernandez-Granda, and Rahul Satija. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nature Biotechnology*, 2023.
4. Maximilian Ilse, Jakub Tomczak, and Max Welling. Attention-based deep multiple instance learning. In *International conference on machine learning*, pages 2127–2136. PMLR, 2018.
5. Caltech-UW TMC Cai Long lcai@ caltech. edu 21 b Shendure Jay 9 Trapnell Cole 9 Lin Shin shinlin@ uw. edu 2 e Jackson Dana 9, UCSD TMC Zhang Kun kzhang@ bioeng. ucsd. edu 15 b Sun Xin 15 Jain Sanjay 24 Hagood James 25 Pryhuber Gloria 26 Kharchenko Peter 8, California Institute of Technology TTD Cai Long lcai@ caltech. edu 21 b Yuan Guo-Cheng 35 Zhu Qian 35 Dries Ruben 35, Harvard TTD Yin Peng peng_yin@ hms. harvard. edu 36 37 b Saka Sinem K. 36 37 Kishi Jocelyn Y. 36 37 Wang Yu 36 37 Goldaracena Isabel 36 37, Purdue TTD Laskin Julia jlaskin@ purdue. edu 10 b Ye DongHye 10 38 Burnum-Johnson Kristin E. 39 Piehowski Paul D. 39 Ansong Charles 39 Zhu Ying 39, Stanford TTD Harbury Pehr harbury@ stanford. edu 11 b Desai Tushar 40 Mulye Jay 11 Chou Peter 11 Nagendran Monica 40, et al. The human body at cellular resolution: the nih human biomolecular atlas program. *Nature*, 574(7777):187–192, 2019.
6. Mohammad Lotfollahi, Sergei Rybakov, Karin Hrovatin, Soroor Hediyeh-Zadeh, Carlos Talavera-López, Alexander V Misharin, and Fabian J Theis. Biologically informed deep learning to query gene programs in single-cell atlases. *Nature Cell Biology*, 25(2):337–350, 2023.
7. Igor Mandric, Tommer Schwarz, Arunabha Majumdar, Kangcheng Hou, Leah Briscoe, Richard Perez, Meena Subramaniam, Christoph Hafemeister, Rahul Satija, Chun Jimmie Ye, et al. Optimized design of single-cell rna sequencing experiments for cell-type-specific eqtl analysis. *Nature communications*, 11(1):5504, 2020.
8. Djamel Nehar-Belaid, Seunghee Hong, Radu Marches, Guo Chen, Mohan Bolisetty, Jeanine Baisch, Lynnette Walters, Marilynn Punaro, Robert J Rossi, Cheng-Han Chung, et al. Mapping systemic lupus erythematosus heterogeneity at the single-cell level. *Nature immunology*, 21(9):1094–1106, 2020.
9. Thomas M Norman, Max A Horlbeck, Joseph M Replogle, Alex Y Ge, Albert Xu, Marco Jost, Luke A Gilbert, and Jonathan S Weissman. Exploring genetic interaction manifolds constructed from rich single-cell phenotypes. *Science*, 365(6455):786–793, 2019.
10. Irene Papatheodorou, Pablo Moreno, Jonathan Manning, Alfonso Muñoz-Pomer Fuentes, Nancy George, Silvie Fexova, Nuno A Fonseca, Anja Füllgrabe, Matthew Green, Ni Huang, et al. Expression atlas update: from tissues to single cells. *Nucleic acids research*, 48(D1):D77–D83, 2020.
11. Stefan Peidli, Tessa D Green, Ciyue Shen, Torsten Gross, Joseph Min, Samuele Garda, Bo Yuan, Linus J Schumacher, Jake P Taylor-King, Debora S Marks, et al. scperturb: harmonized single-cell perturbation data. *Nature Methods*, 21(3):531–540, 2024.
12. Linhao Qu, Manning Wang, Zhijian Song, et al. Bi-directional weakly supervised knowledge distillation for whole slide image classification. *Advances in Neural Information Processing Systems*, 35:15368–15381, 2022.
13. Kanchana Ranasinghe, Muzammal Naseer, Munawar Hayat, Salman Khan, and Fahad Shahbaz Khan. Orthogonal projection loss. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 12333–12343, 2021.
14. Orit Rozenblatt-Rosen, Michael JT Stubbington, Aviv Regev, and Sarah A Teichmann. The human cell atlas: from vision to reality. *Nature*, 550(7677):451–453, 2017.
15. Zhuchen Shao, Hao Bian, Yang Chen, Yifeng Wang, Jian Zhang, Xiangyang Ji, et al. Transmil: Transformer based correlated multiple instance learning for whole slide image classification. *Advances in neural information processing systems*, 34:2136–2147, 2021.
16. Christopher S Smillie, Moshe Biton, Jose Ordovas-Montanes, Keri M Sullivan, Grace Burgin, Daniel B Graham, Rebecca H Herbst, Noga Rogel, Michal Slyper, Julia Waldman, et al. Intra-and inter-cellular rewiring of the human colon during ulcerative colitis. *Cell*, 178(3):714–730, 2019.
17. Yapeng Su, Daniel Chen, Dan Yuan, Christopher Lausted, Jongchan Choi, Chengzhen L Dai, Valentin Voillet, Venkata R Duvvuri, Kelsey Scherler, Pamela Troisch, et al. Multi-omics resolves a sharp disease-state shift between mild and moderate covid-19. *Cell*, 183(6):1479–1495, 2020.
18. Duanchen Sun, Xiangnan Guan, Amy E Moran, Ling-Yun Wu, David Z Qian, Pepper Schedin, Mu-Shui Dai, Alexey V Danilov, Joshi J Alumkal, Andrew C Adey, et al. Identifying phenotype-associated subpopulations by integrating bulk and single-cell sequencing data. *Nature biotechnology*, 40(4):527–538, 2022.

19. Haiyue Wang and Xiaoke Ma. Learning deep features and topological structure of cells for clustering of scrna-sequencing data. *Briefings in Bioinformatics*, 23(3):bbac068, 2022.
20. Haiyue Wang and Xiaoke Ma. Learning discriminative and structural samples for rare cell types with deep generative model. *Briefings in bioinformatics*, 23(5):bbac317, 2022.
21. F Alexander Wolf, Philipp Angerer, and Fabian J Theis. Scanpy: large-scale single-cell gene expression data analysis. *Genome biology*, 19:1–5, 2018.
22. Guangzhi Xiong, Stefan Bekiranov, and Aidong Zhang. Protocell4p: an explainable prototype-based neural network for patient classification using single-cell rna-seq. *Bioinformatics*, 39(8):btad493, 2023.
23. Litao Yang, Deval Mehta, Sidong Liu, Dwarikanath Mahapatra, Antonio Di Ieva, and Zongyuan Ge. Tpmil: Trainable prototype enhanced multiple instance learning for whole slide image classification. *arXiv preprint arXiv:2305.00696*, 2023.
24. Yoshiaki Yasumizu, Daiki Takeuchi, Reo Morimoto, Yusuke Takeshima, Tatsusada Okuno, Makoto Kinoshita, Takayoshi Morita, Yasuhiro Kato, Min Wang, Daisuke Motooka, et al. Single-cell transcriptome landscape of circulating cd4+ t cell populations in autoimmune diseases. *Cell Genomics*, 2024.
25. Carly GK Ziegler, Vincent N Miao, Anna H Owings, Andrew W Navia, Ying Tang, Joshua D Bromley, Peter Lotfy, Meredith Sloan, Hannah Laird, Haley B Williams, et al. Impaired local intrinsic immunity to sars-cov-2 infection in severe covid-19. *Cell*, 184(18):4713–4733, 2021.