Briefings in Bioinformatics, 2026, 27, bbag286
https://doi.org/10.1093/bib/bbag286
Published: 5 June 2026
Problem Solving Protocol

OXFORD

scHILL: deciphering individual-level immune cell heterogeneity with single-cell RNA sequencing data

Yi Wang[1,2,3,4], Hongyu Li[5,6], Lun Li[1,2,3], Yongrong Cao[1,2,3,4], Zhijian Duan[1,2,3,4], Shuhui Song[1,2,3,4,*]

[1]National Genomics Data Center, China National Center for Bioinformation, No. 1 Beichen West Road, Chaoyang District, Beijing 100101, China
[2]Beijing Key Laboratory of Intelligent Governance and Application of Biological Big Data, China National Center for Bioinformation, No. 1 Beichen West Road, Chaoyang District, Beijing 100101, China
[3]Beijing Institute of Genomics, Chinese Academy of Sciences, No. 1 Beichen West Road, Chaoyang District, Beijing 100101, China
[4]University of Chinese Academy of Sciences, No. 1 Yanqihu East Road, Huairou District, Beijing 101408, China
[5]School of Advanced Interdisciplinary Sciences, University of Chinese Academy of Sciences, No. 1 Yanqihu East Road, Huairou District, Beijing 101408, China
[6]Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun East Road, Haidian District, Beijing 100190, China

*Corresponding author. National Genomics Data Center, China National Center for Bioinformation, No. 1 Beichen West Road, Chaoyang District, Beijing 100101, China. E-mail: songshh@big.ac.cn

Abstract

Deep learning frameworks have been developed for interpreting single-cell RNA sequencing (scRNA-seq) data and have demonstrated excellent performance across a range of tasks. However, existing methods remain limited in their ability to characterize heterogeneity at the individual level. To address this gap, we present scHILL, a framework that integrates a masked autoencoder (MAE) with a multilayer perceptron (MLP) to decipher phenotypic heterogeneity arises from immune cell heterogeneity among individuals under specific disease conditions. The MAE, pretrained with data augmentation, enables self-supervised feature learning without labels and effectively mitigates the challenge of limited sample size. The MLP further generates a score for each individual to quantify the functional significance of cells and genes. Across multiple datasets, scHILL outperforms existing methods in phenotype prediction and reveals individual-level immune cell heterogeneity in infectious disease, autoimmune disease, and cancer. scHILL provides a generalizable framework for interpreting individual-level scRNA-seq data, thereby facilitating the future realization of personalized medicine.

Keywords **single-cell RNA sequencing, deep learning, immune cell, individual-level heterogeneity, self-supervised learning, personalized medicine**

Introduction

In recent years, with the development of single-cell RNA sequencing (scRNA-seq) technologies, numerous scRNA-seq datasets have been generated, significantly advancing our understanding of diseases, such as infectious diseases, autoimmune diseases, and cancer. Meanwhile, several databases have been established to integrate these datasets, including the Human Cell Atlas (HCA) [1], CZ CELLxGENE [2], and the Tabula Sapiens [3]. Benefiting from the growth in data volume and inspired by the excellent performance of large language models (LLMs) in natural language understanding [4–7], researchers have developed LLM-based single-cell foundation models, such as scGPT [8], Geneformer [9], scFoundation [10], and GeneCompass [11]. These single-cell foundation models conceptualize each cell as a sentence, with the expressed genes and their corresponding expression levels treated as words. By learning the patterns of gene co-expression and contextual dependencies, these models capture the semantic representations underlying cellular states and identities, and demonstrate outstanding performance across various downstream tasks, including batch integration, cell annotation, and drug sensitivity prediction [12]. However, these models primarily pool cells across individuals and focusing on cellular-level heterogeneity. As a consequence, individual-level heterogeneity in cellular composition is often obscured, for instance, the different constituents of immune cells [13]. If each cell is a sentence, each individual is a book containing trillions of sentences. Therefore, there is an urgent need to utilize scRNA-seq data to decipher heterogeneity at the individual level.

To address this problem, several explainable models have been developed to predict an individual's phenotype from scRNA-seq data, providing new insights into disease progression. PHASE utilizes a self-attention module for cell embedding learning, followed by a pivotal attention-based deep multiple instances learning module that aggregates all single-cell information within a sample to predict

Received: December 29, 2025. **Revised:** March 21, 2026. **Accepted:** May 11, 2026

phenotype [14]. singleDeep first trains deep feed-forward artificial neural networks for each cell type, then uses a weighted voting mechanism to integrate the predictions from each cell type and produce the final phenotype prediction for the sample [15]. CloudPred models each patient's data as a mixture of Gaussians and predicts the phenotype based on the distributions of the means and covariances [16]. Deep sets leverages permutation-invariant functions to fit scRNA-seq data from each sample [17]. These models performed well in supervised phenotype prediction, for instance, coronavirus disease 2019 (COVID-19) disease severity prediction and systemic lupus erythematosus (SLE) diagnosis. However, as these models are based on supervised learning, they primarily characterize differences between patients and healthy controls, failing to reveal heterogeneity among patients, which is critical for advancing personalized medicine [18–21]. Notably, patients assigned the same phenotype label can exhibit markedly different clinical trajectories. For example, among individuals classified as severe cases of COVID-19, some eventually recover, whereas others progress to fatal outcomes [22]. These observations highlight the urgent need for approaches capable of elucidating the heterogeneity among individuals who share the same phenotype label.

In this work, we developed scHILL (scRNA-seq data for deciphering the Heterogeneity at the Individual-Level), a model leveraging masked autoencoder (MAE)-based encoders and multilayer perceptron (MLP)-based classifiers to decipher the phenotypic heterogeneity arises from immune cell heterogeneity among individuals under specific disease conditions [23, 24]. MAE is built upon the Vision Transformer (ViT) [25], which models global dependencies among tokens through self-attention. Therefore, MAE integrates discrete gene expression values into individual-level representations. We also performed data augmentation during pretraining, overcoming the bottleneck of limited sample sizes. scHILL outperformed other methods across several datasets in benchmarking studies, demonstrating its generalization ability. Moreover, scHILL revealed the individual-level immune cell heterogeneity across all three conditions, including infectious disease, autoimmune disease, and cancer. Taken together, scHILL provides new insights into the processing and interpretation of individual-level scRNA-seq data, facilitating the development of personalized medicine, and the discovery of new therapeutic targets.

Materials and methods

Data collection

We collected five publicly available scRNA-seq datasets derived from PBMCs of COVID-19 patients and healthy donors, including the SC4 alliance dataset [22], the COMBAT dataset [26], the Haniffa dataset [27], the Yoshida dataset [28], and the Ye dataset [29] (Table S1). For each dataset, we only retained samples with more than 4000 cells and 10 000 genes. We collected one publicly available scRNA-seq dataset derived from PBMCs of JDM patients and healthy donors, and retained only samples with more than 1000 cells [30]. Ultimately, 15 JDM cases and 5 healthy donors were retained. We collected a publicly available scRNA-seq dataset derived from B cells and plasma cells across 6 organs and retained only samples with more than 224 cells [31]. Ultimately, 24 tumor samples and 19 adjacent non-tumor samples were retained. For each sample across all datasets, we performed quality control using scanpy (version 1.11.4). Mitochondrial genes were annotated based on gene names beginning with the prefix "MT-". Quality control metrics were computed using these mitochondrial genes. Cells with fewer than 200 detected genes or with more than 10% of total counts derived from mitochondrial genes were excluded from subsequent analyses. Furthermore, genes expressed in fewer than three cells were filtered out.

Model pretraining

In each pretraining epoch for COVID-19 cases, we first randomly selected a sample and accessed its expression matrix, $X \in \mathbb{R}^{M \times N}$, where M represented the number of cells and N represented the number of genes. We then sorted X by arranging the columns in descending order according to their mean expression levels, and only the top 2500 genes with the highest mean expression were retained for subsequent analysis. Next, we sorted the rows in descending order by the value in the first column (i.e. the expression level of the gene with the highest mean expression across cells) and finally obtained $X_{sorted} \in \mathbb{R}^{M \times 2500}$. Then, X_{sorted} was randomly cropped to a smaller one, $X_{ramdom} \in \mathbb{R}^{4000 \times 2500}$ with 4000 cells and 2500 genes. The matrix was then split row-wise into three sub-matrices X_1, X_2, and X_3, with 1000, 1000, and 2000 rows, respectively. To fit the input sizes of MAE, these three sub-matrices were projected into matrices with the size 448 × 448 as:

$$X_i A_i = X_{input,i}, i = 1, 2, 3, \tag{1}$$

where A_i was the trainable projection matrix for each X_i and $X_{input,\,i} \in \mathbb{R}^{448 \times 448}$ represented the projected matrix.

To reduce the loss of information caused by projection, we utilized mean square error (MSE) as the loss function to optimize A_i:

$$X_{reconstruction,i} = X_{input,i} A_i{}^{+}, i = 1, 2, 3 \tag{2}$$

$$Projecton\ loss = MSE\left(X_i, X_{reconstruction,i}\right), \tag{3}$$

where $A_i{}^{+}$ represented the pseudoinverse of A_i.

To perform data augmentation, we then randomly accessed a group of coefficients to generate new samples as follows:

$$\sum_{i=1}^{3} a_{j,i} = 1, i = 1, 2, 3, j = 1, \ldots, 17 \tag{4}$$

$$X_j = \sum_{i=1}^{3} a_{j,i} X_{input,i}, \tag{5}$$

where $a_{j,\,i}$ represented the coefficient for $X_{input,\,i}$ in the j-th group.

After that, $X_{input,\,i}$ and X_j were combined into the tensor $I \in \mathbb{R}^{20 \times 448 \times 448}$, representing 20 samples with the size of 448 × 448, as the input of MAE.

For each sample $I_k \in \mathbb{R}^{448 \times 448}$ in I, MAE first divided it into 784 uniform patches of size 16 × 16, then masked 60% of the patches at random. Consequently, I_k was divided into $I_{masked} \in \mathbb{R}^{470 \times 16 \times 16}$ and $I_{unmasked} \in \mathbb{R}^{314 \times 16 \times 16}$. After that, $I_{unmasked}$ was transformed into the latent vector $L_k \in \mathbb{R}^{470 \times 256}$ using the MAE encoder, and all the patches, including masked patches, were reconstructed by the MAE decoder using L_k. The calculation of the reconstruction loss only considered the masked patches, as follows:

$$Loss = MSE\left(I_{masked}, S\left(D\left(E\left(I_{unmasked}\right)\right)\right)\right), \tag{6}$$

where *E(x)*, a six-layer transformer, and *D(x)*, a four-layer transformer, represented the encoder and decoder of MAE, respectively, and *S(x)* represented the selection of masked patches. MAE finally summed

the *Loss* of each sample in I_k and optimized the parameters in *E(x)* and *D(x)*.

The pretraining processes for JDM cases and cancer cases were broadly similar to those for COVID-19, with only minor differences. In both cases, the expression matrix $X \in \mathbb{R}^{M \times N}$ was sorted using the same scheme as for COVID-19, with only one modification: the top 500 and 5000 genes with the highest mean expression were retained in the X_{sorted} of JDM and cancer, respectively. The X_{random} was still randomly cropped from the X_{sorted}, but its size was changed to 224 × 224. Then, we directly took X_{sorted} as I without operating. For $I_k \in \mathbb{R}^{224 \times 224}$, MAE first divided it into 64 uniform patches of size 28 × 28, and then masked 60% of the patches. Consequently, I_k was divided into $I_{masked} \in \mathbb{R}^{38 \times 28 \times 28}$ and $I_{unmasked} \in \mathbb{R}^{26 \times 28 \times 28}$. The loss function was also calculated as in Equation (6), where *E(x)* was a two-layer transformer, and *D(x)* was a one-layer transformer.

Transforming the expression matrices into latent representations

In the COVID-19 case, for the expression matrix $X_{COVID\text{-}19} \in \mathbb{R}^{M \times N}$ per sample, we retained only the 2240 highly variable genes (HVGs), except for the Haniffa dataset (896 HVGs remaining). The HVGs were calculated using the *pp.highly_variable_genes* function in scanpy (version 1.11.4) based on whole-sample data for each dataset. To eliminate cells with lower activity, we sorted the rows by descending *RPS18* expression, a housekeeping gene that regulates translation, and reserved the top 2240 cells. Finally, we got $X_{COVID\text{-}19\text{-}sorted} \in \mathbb{R}^{2240 \times 2240\ (896)}$. Then, $X_{COVID\text{-}19\text{-}sorted}$ was divided into 25 uniform patches (10 for the Haniffa dataset) of size 448 × 448, and we took them as the input $I_{COVID\text{-}19} \in \mathbb{R}^{25\ (10) \times 448 \times 448}$. The pretrained encoder $E_{COVID\text{-}19}$ *(x)* of the MAE with frozen parameters was then utilized to transform I to the latent representation $L_{COVID\text{-}19} \in \mathbb{R}^{25\ (10) \times 784 \times 256}$.

In the JDM case, for the expression matrix $X_{JDM} \in \mathbb{R}^{M \times N}$ of each sample, we firstly retained 224 HVGs. The HVGs were calculated with the *pp.highly_variable_genes* function of scanpy (version 1.11.4) based on each sample. Then, we sorted the rows by variance in descending order and reserved the top 224 cells. Finally, we obtained $X_{JDM\text{-}sorted} \in \mathbb{R}^{224 \times 224}$, which was directly used as the input $I_{JDM} \in \mathbb{R}^{1 \times 224 \times 224}$. The pretrained encoder $E_{JDM}(x)$ of the MAE, with its parameters frozen, was then utilized to transform I_{JDM} to the latent representation $L_{JDM} \in \mathbb{R}^{1 \times 64 \times 64}$.

In the cancer case, for the expression matrix $X_{cancer} \in \mathbb{R}^{M \times N}$ of each sample, we retained 896 HVGs. The HVGs were calculated using the *pp.highly_variable_genes* function of scanpy (version 1.11.4) based on the whole-sample dataset. Then, we sorted the rows in descending order by *FCRL4* expression level, which has been reported as a marker of tumor-enriched B cells, and reserved the top 224 cells. Finally, we got $X_{cancer\text{-}sorted} \in \mathbb{R}^{224 \times 896}$. Then, $X_{cancer\text{-}sorted}$ was divided into 3 uniform patches of size 224 × 224, and we took them as the input $I_{cancer} \in \mathbb{R}^{3 \times 224 \times 224}$. The pretrained encoder $E_{cancer}(x)$ of the MAE with frozen parameters was then utilized to transform I_{cancer} to the latent representation $L_{cancer} \in \mathbb{R}^{3 \times 64 \times 64}$.

Model fine-tuning for phenotype prediction

In the COVID-19 cases, there were three label types: healthy control, severe COVID-19, and mild COVID-19. Fine-tuning was performed with two MLP-based classifiers: one for predicting health or disease, and the other for predicting disease severity. Meanwhile, each dataset was fine-tuned separately. The MLPs were constructed from fully connected neural networks with one input layer, one output layer, and three latent layers.

We first used $Classifier_1$ to determine whether the sample was from a COVID-19 patient or a healthy control. If the sample was considered a patient, we used $Classifier_2$ to determine whether it was from a severe patient or a mild patient. We utilized cross-entropy loss to calculate the loss of phenotype prediction as follows:

$$z = MLPi\left(L_{COVID-19,i,j}\right) \tag{7}$$

$$CE = -\log\left(\frac{e^{z_R}}{\sum_{n=1}^{N} e^{z_n}}\right), \tag{8}$$

where z denoted the result of prediction; $MLP_i(x)$ denoted the i-th MLP-based classifier, including $Classifier_{i,1}$ and $Classifier_{i,2}$ for the i-th training set; $L_{COVID\text{-}19,\ i,\ j}$ denoted the latent representation of the j-th sample in the i-th training set; CE denoted the cross-entropy loss; N denoted the types of labels; z_n denoted the probability of the sample for the *n-th* class; and z_R denoted the probability of the sample for the real label R.

In the JDM case, there were two label types: healthy controls and JDM. Fine-tuning was performed with a single MLP classifier, with one input and one output layer. The process of computing the prediction loss followed Equations (7) and (8), except that the input to Equation (7) was changed to $L_{JDM,\ i,\ j}$.

Benchmarking the performance of phenotype prediction

For the five COVID-19 datasets and one JDM dataset used for phenotype prediction, we split them into training, validation, and test sets with proportions of 60%, 20%, and 20%, respectively. We benchmarked scHILL against PHASE [14], singleDeep [15], CloudPred [16], and DeepSets [17]. The methods for data preprocessing, training, validation, and testing were consistent with the previously reported schemes for each model.

Model fine-tuning for score generation

In the COVID-19 case, we utilized five-fold cross-validation to generate scores for each patient in the SC4 alliance dataset. For each fold, we fine-tuned scHILL for 500 epochs and retained the model with the highest MCC. We then utilized the saved model to generate scores for patients in the testing set. The softmax function was applied to transform the outputs of $Classifier_2$ into 2 probabilistic values that sum to 1, representing the probabilities of mild and severe COVID-19 cases, respectively. Finally, the scores were calculated as follows:

$$Score = 1 \times P_{mild} + 2 \times P_{severe}, \tag{9}$$

where P_{mild} and P_{severe} denoted the probability of COVID-19 mild cases and COVID-19 severe cases, respectively. As $P_{mild} \in [0, 1]$, $P_{severe} \in [0, 1]$, and their sum equaled one, it followed that $Score \in [1, 2]$, with values near 1 corresponding to the mild cases and those near 2 to the severe cases.

In the JDM case, the fine-tuning method was the same as for COVID-19. In each fold, we utilized the saved model to generate scores for patients. The softmax function was applied to transform the model

outputs into 2 probabilistic values, P_{health} and P_{JDM}, which summed to 1 and represented the probabilities of healthy controls and JDM patients, respectively. For each JDM patient, P_{JDM} was directly used as the *Score*.

For both the COVID-19 and JDM cohorts, we performed the above processes five times, and the average of the resulting scores was used as the final score for each patient.

The correlation coefficients and regression coefficients between scHILL scores and gene expression levels was calculated with the *pearsonr()* and LinearRegression() function of Python (version 3.12), respectively. For each individual, we used the mean expression level across all cells as the individual's gene expression level. Genes with $|\text{regression coefficients}| > 0.5$ and $|\text{Pearson's r}| > 0.5$ were designated as high-impact genes.

Single-cell analysis

All single-cell analyses in this study were performed in RStudio (version 2023.6.2.561) based on R (version 4.3.1). For each dataset, we first used the *log1p* function to transform the expression matrices, and then used the cell type annotations from the original studies for subsequent analyses.

Pseudo-time analysis was performed by monocle (version 2.28.0). Genes with lower expression levels were eliminated from the pseudo-time analysis by the *dispersionTable* function with the condition "mean_expression >= 0.0001 & dispersion_empirical >= 1 * dispersion_fit". Dimensional reduction was performed with the *reduceDimension* function with parameters "max_components = 2, method = "DDRTree".

Differential gene expression analysis was performed by the *FindAllMarkers* function of Seurat (version 5.3.0), with the parameters "only.pos = TRUE, min.pct = 0.25, logfc.threshold = 0.25".

Cytotoxicity scores of $CD8^+$ T cells in the COVID-19 case were calculated with the *AddModuleScore* function of Seurat. Twelve previously reported cytotoxicity-associated genes [32], including *PRF1*, *IFNG*, *GNLY*, *NKG7*, *GZMB*, *GZMA*, *GZMH*, *KLRK1*, *KLRB1*, *KLRD1*, *CTSW*, and *CST7*, were used to define the cytotoxicity score.

Results

Overview of scHILL

The scHILL consists of two parts: an MAE-based encoder that transforms expression matrices into a latent space and an MLP-based classifier (Fig. 1a). These two parts are trained during the pretraining and fine-tuning stages, respectively. The pretraining stage is self-supervised, requiring neither cell-type annotations nor phenotype labels, and uses only single-cell expression matrices as input, enabling scHILL to learn the characteristics of each individual adequately. Meanwhile, this also allows scHILL to discover new groups beyond those known labels. Because single-cell expression matrices are large, scHILL first crops them into smaller fixed-size matrices. One single-cell expression matrix can be split into multiple smaller matrices, enabling our scheme to process the expression matrix and overcoming the bottleneck that small sample sizes impose on self-supervised pretraining at the individual level. After that, the MAE randomly masks the cropped matrices. The masked matrices are taken as inputs and then transformed into the latent space. The encoder of MAE is based on ViT, which is capable to capture global dependencies through self-attention. Therefore, the encoder integrates discrete gene expression values into individual-level representations. Finally, the decoder recovers the matrices, and the reconstruction losses between outputs and inputs are used to optimize the model parameters. The masking and reconstruction mechanism randomly removes a substantial proportion of input features and requires MAE to reconstruct them using only the visible subset. This forces the encoder to rely on global signals rather than noises and artifacts.

In the fine-tuning stage, the pretrained encoder is frozen, and only the MLP-based classifier is trained. The classifier predicts phenotypes and generates scores for each individual, facilitating the identification of cells and genes that contribute to individual-level heterogeneity. Moreover, because individuals with the same phenotype label may have variable scores, scHILL not only distinguishes distinct phenotype groups (e.g. disease versus health) but also reveals heterogeneity within each phenotype group. Taken together, scHILL provides a new framework for individual-level interpretation of scRNA-seq data and offers a complementary dimension to disease assessments (Fig. 1b).

Robust phenotype prediction and biological interpretability of scHILL

To evaluate the feasibility of scHILL, we assessed its performance in phenotype prediction, using three publicly available scRNA-seq datasets derived from peripheral blood mononuclear cells (PBMCs) of COVID-19 patients and healthy donors, namely the SC4 alliance dataset [22], the COMBAT dataset [26], and the Haniffa dataset [27]. The samples are categorized into three clinical groups (Table S1), including 39 healthy controls, 74 mild COVID-19 cases, and 79 severe COVID-19 cases. We first performed self-supervised pretraining on the combined dataset, then fine-tuned on each dataset. The performance of scHILL was then benchmarked against PHASE, singleDeep, CloudPred, and Deep Sets, using precision, recall, and F1-score as evaluation metrics. The result of benchmarking exhibited that scHILL achieved the best performance on the SC4 alliance dataset and COMBAT dataset, and showed comparable performance to PHASE on the Haniffa dataset (Fig. 2a–c). Since the classifier of scHILL used in the fine-tuning stage was merely a simple four-layer MLP, these results demonstrated that the pretrained encoder possessed a strong capability of feature extraction and representation. To further assess scHILL's generalization capability, we collected two additional datasets not included in the pretraining stage [28, 29]. We froze the parameters of the pretrained encoder and fine-tuned the classifier on each dataset separately. scHILL consistently outperformed all other models across these datasets, demonstrating its robust performance across diverse single-cell cohorts (Fig. S1).

To further investigate the biological interpretability of scHILL, we next examined the high-impact genes identified by the model. scHILL identified 696 high-impact genes with substantial contributions to phenotype prediction in the SC4 alliance dataset. Given that age also affects immune cells [33], we then performed a correlation analysis between age and gene expression levels. Consequently, only 24 genes showed strong correlations with ages ($|\text{Pearson's r}| > 0.5$), demonstrating that the scHILL scores more faithfully reflected intrinsic transcriptional programs. We then performed Gene Ontology (GO) enrichment analysis on the high-impact genes and found that they were primarily enriched in T cell-related pathways (Fig. S2a), which was in line with the previous findings that T cells were abundant and variable in the PBMCs in COVID-19 patients [22, 32–35]. To further verify the robustness of scHILL, we conducted the same analysis on

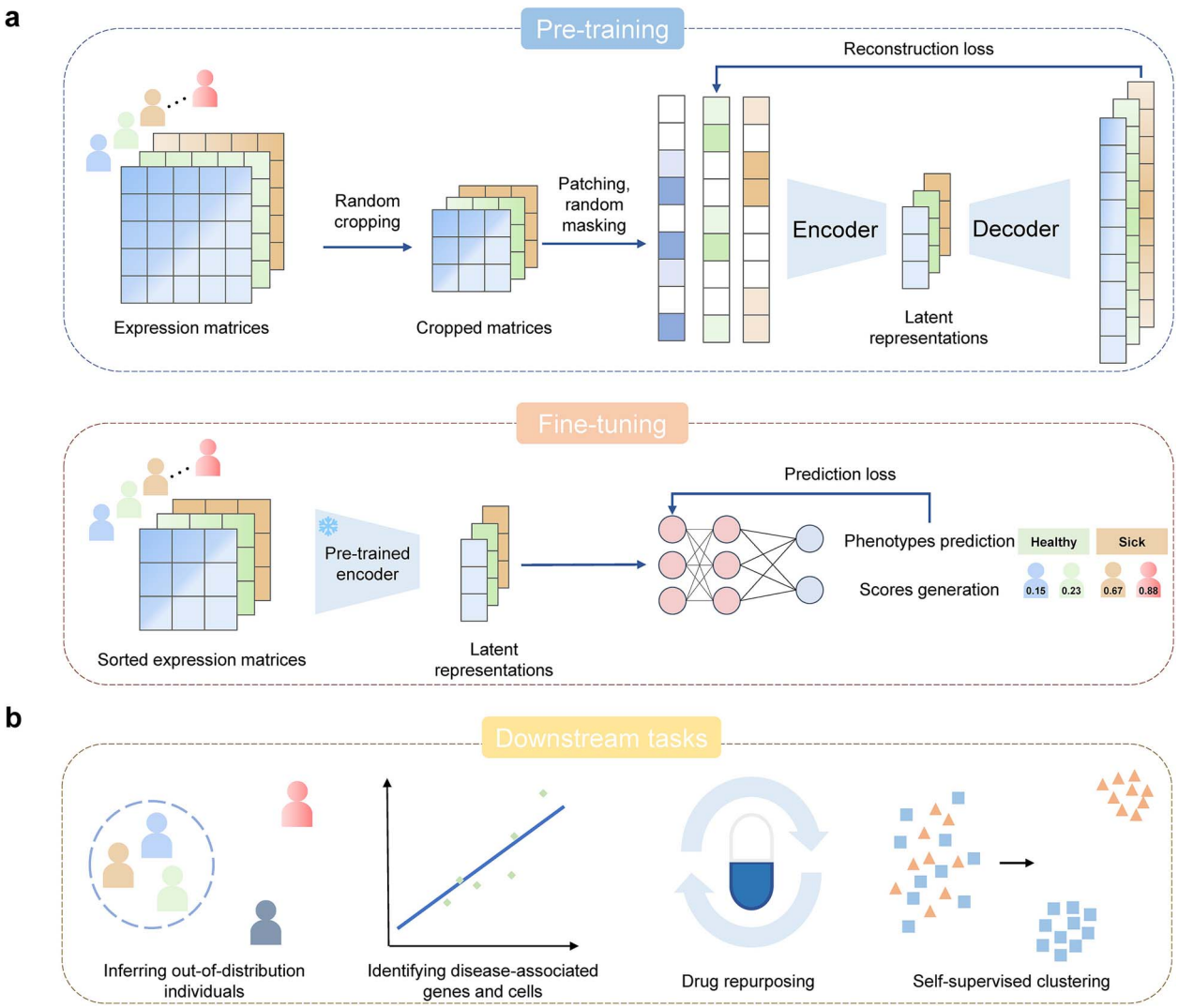

Figure 1 Overview of the scHILL framework and workflow. (a) The training process of scHILL. The first step is the self-supervised pretraining of MAE, in which the encoder transforms the expression matrices into the latent space and then the decoder reconstructs the masked patches, aiming for improving the capability of the encoder to capture the meaningful features in the single-cell expression matrices. The second step is the supervised fine-tuning of MLP-based classifier, in which the MLP predicts the phenotypes of samples and generates scores for subsequent analyses. (b) Downstream tasks enabled by scHILL.

the COMBAT dataset. Ultimately, a total of 364 genes were identified as high-impact genes, and none showed a strong correlation with age ($|\text{Pearson's r}| > 0.5$). Moreover, GO enrichment analysis also revealed that the high-impact genes were significantly enriched in T cell-related pathways (Fig. S2b). Taken together, scHILL exhibited excellent performance in phenotype prediction and generated biologically interpretable scores.

scHILL enables finer stratification of patients and reveals distinct immune subgroups

Finer stratification of patients is a fundamental prerequisite for the realization of personalized medicine. At the cellular level, out-of-distribution (OOD) cells deviate from established cellular paradigms due to their distinct molecular characteristics, reflecting heterogeneity across disease states [36–38]. Extending this concept to the individual level, we hypothesized that scRNA-seq data can also be used to identify OOD patients for finer stratification. To this end, we assumed that mild and severe COVID-19 patients each followed a distinct distribution of scHILL scores and fitted these distributions using Gaussian mixture models (GMMs) in the SC4 alliance dataset. The result revealed a bimodal distribution corresponding to mild patients (mean = 1.36, standard deviation = 0.23) and severe patients (mean = 1.82, standard deviation = 0.16) (Fig. 3a). Subsequently, mild patients with scHILL scores greater than 1.59 (1.36 + 0.23) were designated as OOD-mild patients. Meanwhile, severe patients with scHILL scores lower than 1.66 (1.82–0.16) were designated as OOD-severe patients. In total, 3 out of 14 mild patients and 6 out of 26 severe patients were in the OOD-mild and OOD-severe groups, respectively. Notably, only 1 OOD-severe patient, aged 92—the oldest among all participants—died, whereas 5 of the 20 severe patients succumbed (odds ratio = 1.67), suggesting that our classification scheme effectively delineated clinically relevant heterogeneity.

To assess the biological significance of our finer stratification in COVID-19 patients, we first compared the proportions of different cell types. Across all groups, $CD8^+$ T cells, $CD4^+$ T cells, and monocytes

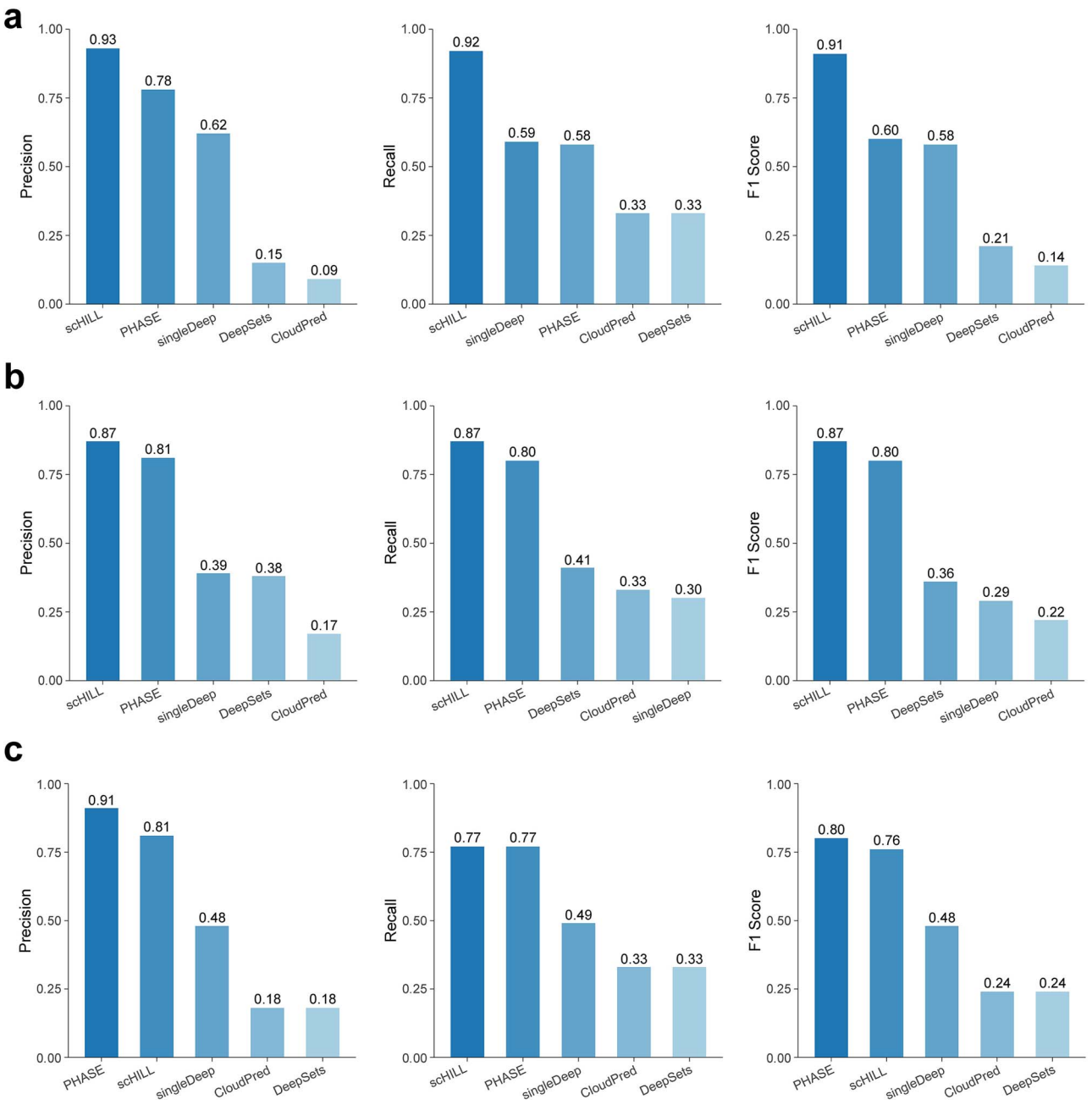

Figure 2 Performance of scHILL and other models on phenotype prediction. (a–c) The precision, recall, and F1-score of each model in COVID-19 severity prediction (totally three phenotypes, including heath, mild, and severe) with SC4 alliance dataset (a), COMBAT dataset (b), and Haniffa dataset (c), respectively.

were consistently the three most abundant cell types (Fig. S3), reflecting a shared feature of the immune landscape. Nevertheless, the proportions of cell types differed significantly between groups ($F = 2.144$, $P = .038$), motivating a more detailed examination of specific cell populations. In cell-mediated immunity, previous reports indicated that $CD8^+$ T cells were significantly decreased in PBMCs from severe patients compared with those from mild patients [22, 26, 27, 32]. However, in our analysis, although the median proportion of $CD8^+$ T cells gradually decreased across groups, the difference between the mild group and OOD-severe group was not statistically significant (Fig. 3b). This phenomenon was also observed in several subtypes of $CD8^+$ T cells, such as T_CD8_c04-*COTL1*, T_CD8_c05-*ZNF683*, T_CD8_c07-*TYROBP*, T_CD8_c08-*IL2RB*, T_CD8_c12-*MKI67*-*TYROBP*, and T_CD8_c13-*HAVCR2* (Fig. S4). We further calculated cytotoxicity scores for $CD8^+$ T cells to evaluate their capacity to eliminate infected cells and found that the mild, OOD-mild, and OOD-severe groups showed comparable distributions of scores. In contrast, the severe group had significantly lower cytotoxicity scores (Fig. 3c). For humoral immunity, we found that the OOD-mild group exhibited a significantly higher proportion of B cells than the mild group. This feature has previously been reported only in severe COVID-19 cases (Fig. 3d) [22, 26, 27, 32]. In contrast, the OOD-severe

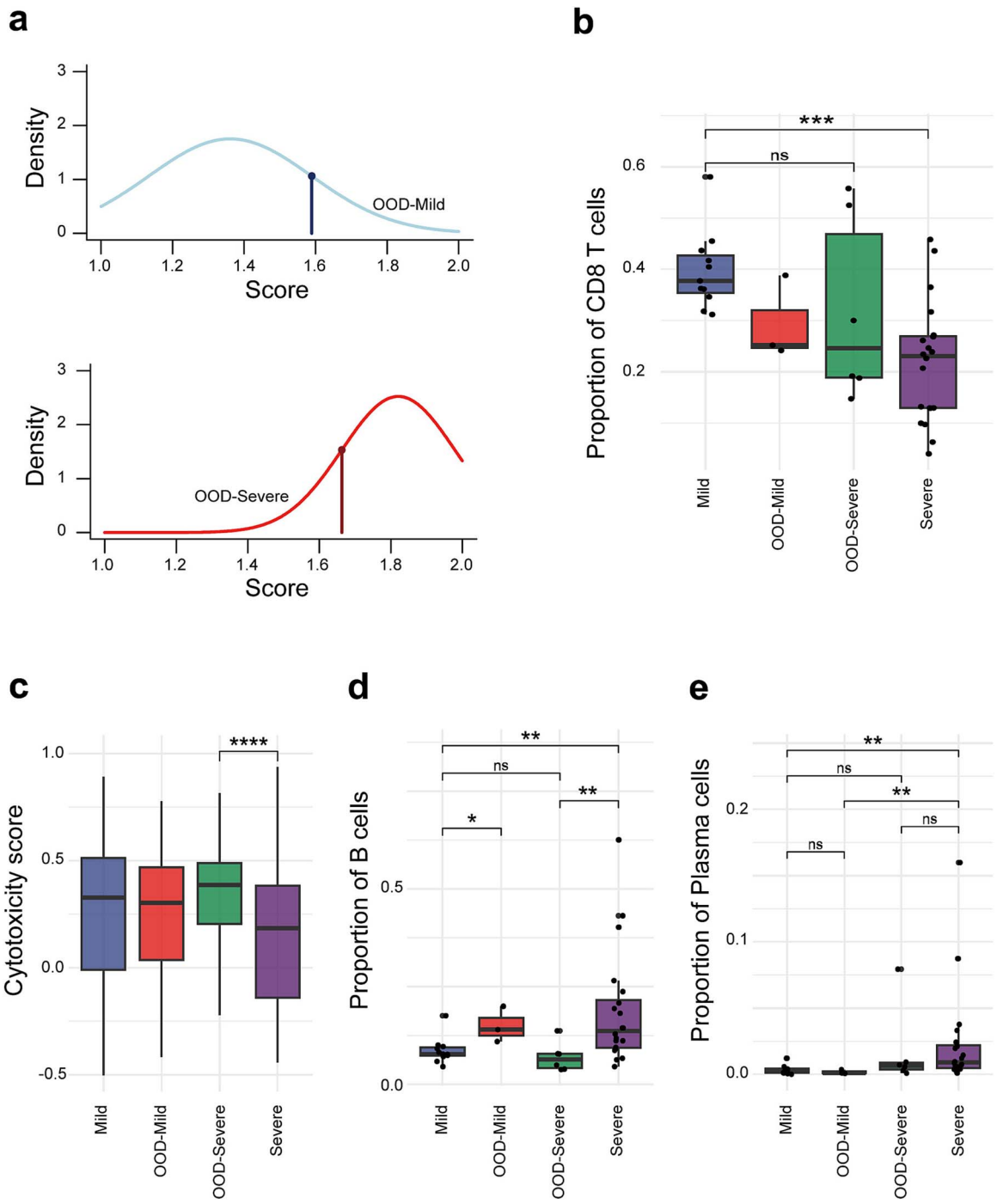

Figure 3 Finer stratification of COVID-19 patients and the immune characteristics of two new groups. (a) The fitted distribution of scHILL scores for mild and severe COVID-19 patients. The vertical lines denote the boundary of OOD-patients (1.59 for OOD-mild and 1.66 for OOD-severe, respectively). (b) Proportion of $CD8^+$ T cells in each group. (c) Cytotoxicity scores of $CD8^+$ T cells in each group. (d) Proportion of B cells in each group. (e) Proportion of plasma cells in each group. Components of the box plot are: center line, median value; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range. The Mann–Whitney U test was used to calculate the statistical significance (*P*-value, with ns > 0.05, * < 0.05, ** < 0.01, *** < 0.001, and **** < 0.0001).

group showed a proportion of B cells similar to that of the mild group, and this was also observed across B cell subtypes (Figs 3d and S5). Subsequent analysis of plasma cells revealed no significant difference between the OOD-mild and mild groups (Fig. 3e), suggesting that the elevated B cells in the OOD-mild group did not differentiate into plasma cells. To verify the robustness of our findings, we then performed the same analyses using the COMBAT dataset and found the consistent characteristics of $CD8^+$ T cells and B cells (Fig. S6). These results indicated that scHILL enabled finer stratification of patients and revealed distinct immune characteristics in both cell-mediated and humoral immunity. Specifically, OOD-severe patients exhibited cell-mediated immune profiles similar to mild patients, whereas OOD-mild patients resembled severe patients in humoral immunity.

scHILL reveals potential therapeutic targets and facilitates drug repurposing

Juvenile dermatomyositis (JDM) is an autoimmune disease whose treatment primarily relies on broad-spectrum immunosuppressants, such as corticosteroids and methotrexate [39]. Therefore, identifying precise therapeutic targets for JDM is of great importance. To achieve this, we pretrained and fine-tuned scHILL on a scRNA-seq dataset derived from PBMCs of JDM patients and healthy donors, containing 15 disease cases and 5 healthy controls [30]. scHILL achieved the best performance in the benchmark against other methods (Fig. 4a), demonstrating the generalizability of scHILL under small sample size conditions. Meanwhile, scHILL identified 31 high-impact genes (Table S2), among which 20 genes exhibited negative contributions and 11 genes exhibited positive contributions to the phenotype prediction, respectively. Notably, several NK cell markers were found among the negatively contributing genes, such as *FCGR3A*, *ITGAL*, *KLRD1*, *KLRF1*, and *FGFBP2*. We subsequently inspected the correlation between the proportion of NK cells and scHILL scores, and found a significantly negative correlation (Pearson's $r = -0.55$, $P = .0339$) (Fig. 4b). These results suggested that, although scHILL was not trained with explicit cell-type information, it could effectively capture changes in cell composition. Our finding was consistent with a previous report that disease activity scores of JDM patients were negatively correlated with the proportion of NK cells (Pearson's $r = -0.324$, $P = .0002$) [40]. Among the 11 positively contributing genes, *CEMIP2* (Regression coefficient = 1.83, Pearson's $r = 0.73$, $P = .00196$) and *PDE3B* (Regression coefficient = 2.62, Pearson's $r = 0.64$, $P = .0106$) were reported as the therapeutic targets of autoimmune disease and heart disease (Table S2) [41–43]. Moreover, PDE3B inhibitors, such as milrinone and cilostazol, had already been approved for clinical use [43]. To evaluate their potential for drug repurposing in JDM, we performed single-cell analyses to uncover the role of *PDE3B* in JDM. We found that the expression of *PDE3B* was most prevalent in naive $CD8^+$ T cells in both JDM patients and healthy controls, in which 43.5% and 41% of the cells expressed *PDE3B*, respectively, followed by naive $CD4^+$ T cells, with 25.1% and 22.6% of the cells expressing *PDE3B* (Fig. S7). Meanwhile, when considering the absolute counts of $PDE3B^+$ cells, naive $CD4^+$ and $CD8^+$ T cells were also the two predominant populations in both JDM patients and healthy controls, with naive $CD4^+$ T cells being slightly more abundant (5615 versus 4396 cells in JDM patients; 1743 versus 1197 cells in healthy controls) (Fig. S8). Furthermore, the expression levels of *PDE3B* in both naive $CD8^+$ $PDE3B^+$ and naive $CD4^+$ $PDE3B^+$ T cells differed significantly between JDM patients and healthy controls (Fig. 4c), highlighting the potential involvement of *PDE3B* in JDM pathogenesis within these two cell types. Taken together, scHILL reconstructed the deficiency of NK cells in JDM patients and uncovered *CEMIP2* and *PDE3B*, which have been used in other pharmaceuticals, as potential therapeutic targets for JDM.

scHILL identifies cross-tissue characteristics of normal adjacent tissues

Previous studies have shown that normal adjacent tissues (NATs) are molecularly distinct from both tumor and healthy tissues [44, 45], and NATs can be further classified into different subtypes in specific cancer types [46–48]. To reveal the cross-tissue characteristics of NATs, we pretrained scHILL on a publicly available scRNA-seq dataset containing B cells and plasma cells from six organs [31]: kidney, breast, uterus, thyroid, pancreas, and esophagus. Except for the pancreas, which was sampled only from tumor tissues, other tissues were sampled from both tumor tissues and NATs, totaling 24 tumor samples and 19 NATs. After self-supervised pretraining, single-cell expression matrices of NATs were transformed into latent representations and visualized using uniform manifold approximation and projection (UMAP). In the latent space, NATs clearly segregated into two distinct clusters (Fig. 5a), whereas this separation was not observed before transformation (Fig. S9a), where samples remained intermixed. We also performed principal component analysis (PCA) and k-nearest neighbors (kNN) on the untransformed expression matrices, but neither method revealed a clear separation (Fig. S9b and c). Taken together, these results demonstrated that the self-supervised framework of scHILL learned more informative representations, enabling cross-tissue discrimination of NATs. We then analyzed the difference between these two clusters. Since cluster A contained all the uterus samples and cluster B contained all the kidney samples, we first compared the expression patterns of the uterus and the kidney. Consequently, we identified several antibody-related genes as markers of the uterus, including *IGHG3*, *IGHG2*, *IGLV3-1*, and *IGLL5* (Table S3). Given that antibody-related genes were mainly expressed in plasma cells, we examined whether the two clusters differed in plasma cell proportions. Ultimately, samples belonging to cluster A exhibited significantly higher proportion of plasma cells (Fig. 5b). Consistently, thyroid and esophagus samples that were classified into cluster A showed higher proportion of plasma cells than their counterparts in cluster B (Fig. 5c). Subsequent pseudo-time analysis revealed the difference in cell differentiation between the two clusters, with cluster A showed continuously B cells differentiation. In contrast, cluster B showed a gap between B cells and plasma cells (Fig. 5d). Taken together, the self-supervised clustering results generated by scHILL suggested that, in some cases, B cells in NATs might be continuously stimulated by the tumor microenvironment and subsequently differentiated into plasma cells.

Conclusion and discussion

In this study, we developed scHILL, a deep learning framework for deciphering phenotypic heterogeneity arises from immune cell heterogeneity among individuals under specific disease conditions. scHILL is comprised of an MAE-based encoder and an MLP-based classifier. Compared with autoencoder-based models such as scGDC, scLDS2, and scAMAC [49–51], MAE is based on a ViT rather than a fully connected network (FCN). The MAE-based encoder that transforms expression matrices into latent representations and the MLP-based classifier that predicts phenotype and generates scHILL scores for subsequent analysis, which are trained separately in the pretraining and fine-tuning stages. The framework of scHILL has following advantages. Firstly, scHILL effectively learns the characteristics of each sample without labels, thereby enhancing its ability to capture subtle differences among individuals with the same label. Second, the random cropping scheme overcomes the bottleneck of small sample sizes. Notably, although the single-cell expression matrices were cropped to smaller sizes than the original matrices, the retained information remained sufficient for effective model training and phenotype prediction (Tables S4–S6). One possible explanation is the high sparsity of single-cell expression matrices. Our analysis

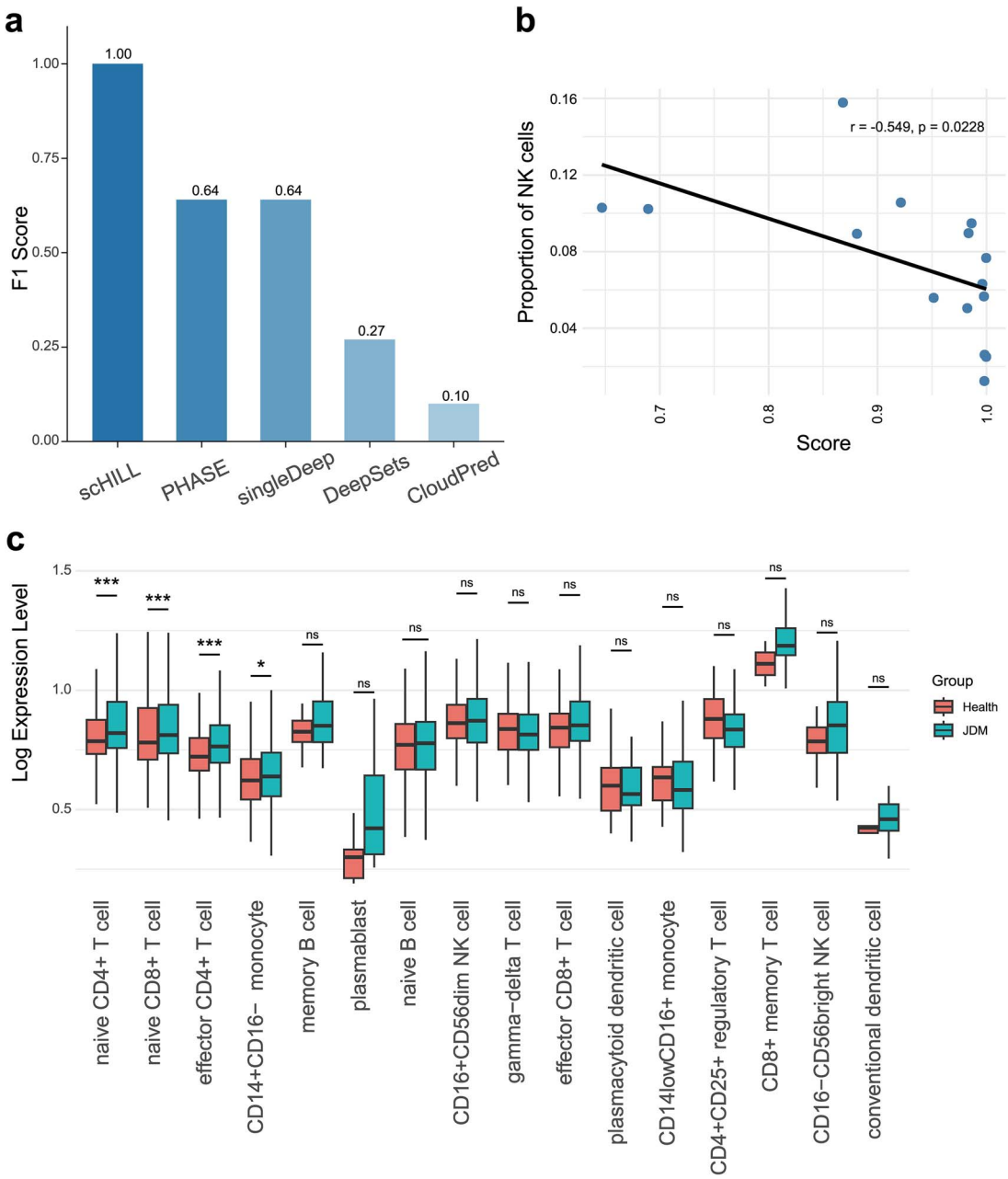

Figure 4 *PDE3B* is a potential therapeutic target of JDM. (a) F1-scores of different models for JDM prediction. (b) Correlation between scHILL scores and proportion of NK cells. (c) Comparison of the expression levels of *PDE3B* in each *PDE3B*$^{+}$ cell type between JDM patients and healthy controls. Cell types are ranked by adjusted *P*-value and displayed from left to right. The Mann–Whitney U test was used to calculate the statistical significance (*P*-value, with ns $>$ 0.05, * $<$ 0.05, ** $<$ 0.01, *** $<$ 0.001, and **** $<$ 0.0001).

further revealed that the cropped matrices contained relatively higher expression levels and a lower proportion of zero values compared with the removed portions of the matrices (Fig. S10). Third, MAE is capable to capture global interdependencies through self-attention, and its masking and reconstruction mechanism forces the encoder to rely on global signals rather than noises and artifacts. Through benchmarking on several PBMC datasets, scHILL exhibited excellent performance in phenotype prediction and generated clinically meaningful scores for samples. scHILL also uncovered the characteristics of plasma cells in solid tissues, demonstrating its flexibility and scalability.

Previous reports have revealed the differences in immune cells between COVID-19 mild and severe cases [22, 26–28, 32, 52, 53]. It is noteworthy that this comparison cannot distinguish among patients with the same clinical severity. For instance, some COVID-19 severe patients recovered while others died, and some COVID-19 mild patients recovered fast while others recovered slowly and even suffered from long-COVID [54, 55]. To fill this gap, inspired by OOD cells, we utilized scHILL to identify OOD groups. Firstly, we benchmarked scHILL against other methods for COVID-19 severity prediction and found that it outperformed them, demonstrating its ability to capture immune features. Then, scHILL generated scores

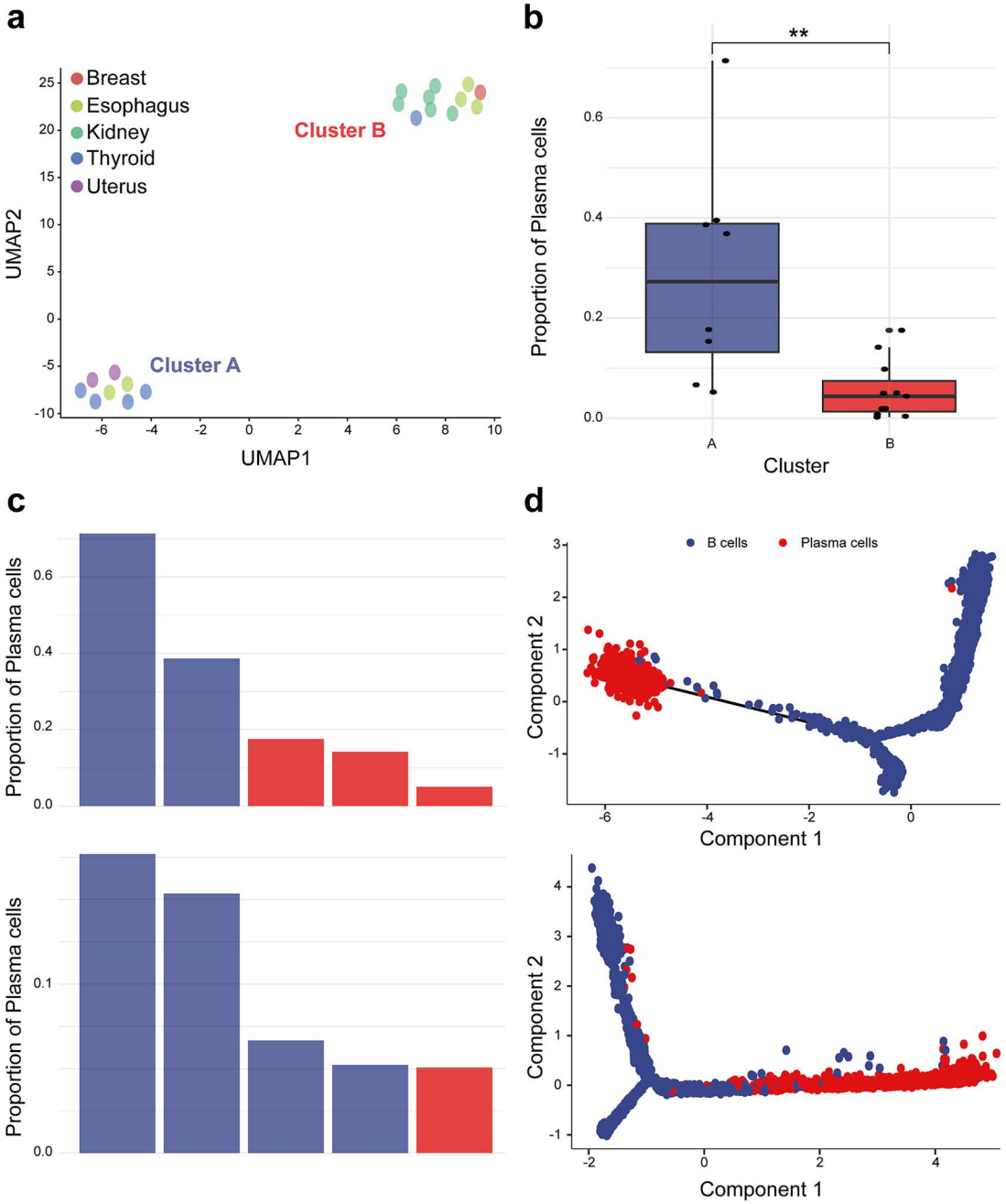

Figure 5 Cross-tissue heterogeneity of plasma cells in NATs. (a) UMAP visualization of the latent representations of each sample. (b) Comparison of the proportion of plasma cells between cluster A and cluster B. Components of the box plot are: center line, median value; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range. The Mann–Whitney U test was used to calculate the statistical significance (*P*-value, with ns > 0.05, * < 0.05, ** < 0.01, *** < 0.001, and **** < 0.0001). (c) The proportion of plasma cells among the same tissues within different clusters. (d) Trajectories of B cells and plasma cells in cluster A and cluster B.

for each individual and identified OOD groups. Subsequent analysis showed that the OOD-severe group indeed had a proportion of T cells and T cell effect scores similar to those of the mild group, indicating a stronger ability to eliminate infected cells and verifying the biological meaning of scHILL scores. Meanwhile, the OOD-mild group indeed had a similar proportion of B cells as the severe group, further demonstrating the effectiveness of scHILL in revealing heterogeneity among individuals with the same phenotype. A previous report has identified individual-level signatures of SLE and finely stratified SLE patients into seven groups based on blood transcriptional signatures derived from microarray data [56]. Our results showed that scRNA-seq data can also be used for more refined stratification.

JDM is an autoimmune disease whose treatment relies on broad-spectrum immunosuppressants, such as corticosteroids and methotrexate [39]. Therefore, it is essential to detect the therapeutic targets for JDM. For this purpose, we utilized scHILL to generate scores for JDM patients. The scHILL scores reproduced the lack of NK cells in JDM patients and the negative correlation between

clinical scores and NK cell proportions, indicating the reliability of our method. We also found that *CEMIP2* and *PDE3B* showed a strongly positive contribution to scHILL scores, highlighting their roles in JDM progression. Further analysis of *PDE3B* revealed its enrichment in naive T cells and higher expression in JDM patients, suggesting that *PDE3B* may contribute to the development of T cell-driven autoimmune disorders. We also found *IL6ST*, the gene had been previously reported to contribute to tumor promotion and mediate rheumatoid arthritis [57, 58], was highly expressed in $CD8^+$ $PDE3B^+$ naive T cells (Fig. S11, Tables S7 and S8), further suggesting that these cells might play a role in inflammatory signaling pathways relevant to JDM. Drugs targeting *CEMIP2* and *PDE3B* are already available [41–43], and they may be evaluated in future clinical trials for the treatment of JDM. In the clinical management of autoimmune diseases, disease activity scores, which are discrete and based on clinical manifestations, have been widely used [59–61]. The scHILL scores are continuous and derived from single-cell profiles, offering a complementary dimension to disease assessments.

NATs are involved in the tumor microenvironment and play a vital role in tumorigenesis and metastasis [62]. Previous reports have shown that NATs are distinct from both healthy and tumor tissues [44, 45], but the cross-tissue patterns of NATs remain unclear. In the cross-tissue dataset, scHILL divided NATs into two clusters: one showed enrichment for plasma cells and continuous B cell differentiation, while the other showed the opposite. NATs from the thyroid and esophagus were present in both clusters, demonstrating their intrinsic heterogeneity. The proportion of plasma cells varied in the NATs of non-small cell lung cancer (NSCLC), and plasma cells were associated with the efficacy of programmed cell death-1 PD-L1 blockade [63]. Our findings suggest that this association might extend across different tissue types. Given that plasma cells represent a double-edged sword in tumorigenesis, as antibodies secreted by them facilitate antitumor immunity by killing cancer cells, whereas some plasma cells are immunosuppressive, which can impede the effectiveness of chemotherapy [64], further studies with larger cohorts are required to clarify whether the expanded plasma cells confer beneficial or detrimental outcomes.

As scRNA-seq datasets derived from various diseases are continuously being released, scHILL is expected to play a vital role in uncovering the heterogeneity among individuals. Furthermore, its self-supervised pretraining scheme is well-suited to conditions with unknown labels, such as the enrichment or deficiency of plasma cells in the cancer dataset in this study, enabling the discovery of new insights into disease progression. In the next phase, we aim to apply scHILL not only to disease cases but also to healthy individuals to uncover age- and ethnicity-associated differences.

Key Points

- scHILL utilizes masked autoencoder to perform self-supervised representation learning at the individual level.
- scHILL supports robust phenotype prediction and finer stratification, revealing individual-level heterogeneity.
- scHILL reveals clinically relevant immune heterogeneity, highlighting its utility for personalized medicine.

Acknowledgements

We thank the members of the China National Center for Bioinformation for providing the computing resource. We thank Dongmei Tian, Xue Bai, Hong Luo, and Fei Yang for helpful discussions.

Author contributions

Shuhui Song (Supervision, Funding acquisition, Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing–original draft, Writing– review & editing), Yi Wang (Conceptualization, Data curation, Formal analysis, Methodology, Software, Visualization, Writing–original draft, Writing–review & editing), Hongyu Li (Methodology, Software), Lun Li (Data curation, Methodology), Yongrong Cao (Methodology, Software), Zhijian Duan (Methodology, Software).

Supplementary material

Supplementary material is available at *Briefings in Bioinformatics* online.

Conflicts of interest

The authors declare no competing interests.

Funding

This work was supported by the National Natural Science Foundation of China (92374201).

Data availability

All of the datasets analyzed in the current study are publicly available. The COVID-19 and JDM datasets are available in CZ CELLxGENE under the following links: the SC4 alliance dataset (https://cellxgene.cziscience.com/collections/0a839c4b-10d0-4d64-9272-684c49a2c8ba), the COMBAT dataset (https://cellxgene.cziscience.com/collections/8f126edf-5405-4731-8374-b5ce11f53e82), the Haniffa dataset (https://cellxgene.cziscience.com/collections/ddfad306-714d-4cc0-9985-d9072820c530), the Ye dataset (https://cellxgene.cziscience.com/collections/7d7cabfd-1d1f-40af-96b7-26a0825a306d), the Yoshida dataset (https://cellxgene.cziscience.com/collections/03f821b4-87be-4ff4-b65a-b5fc00061da7), and the JDM dataset (https://cellxgene.cziscience.com/collections/c672834e-c3e3-49cb-81a5-4c844be4a975). The cancer dataset is available in Gene Expression Omnibus (GEO) under accession number GSE233236 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE233236).

Code availability

The source code of scHILL is written in Python and available at GitHub (https://github.com/wy0411/scHILL).

References

1. Regev A, Teichmann SA, Lander ES *et al*. The human cell atlas. *eLife* 2017;**6**:e27041. https://doi.org/10.7554/eLife.27041

2. Program CCS, Abdulla S, Aevermann B *et al.* CZ CELLxGENE discover: a single-cell data platform for scalable exploration, analysis and modeling of aggregated data. *Nucleic Acids Res* 2025;**53**:D886–900. https://doi.org/10.1093/nar/gkae1142
3. Consortium* TTS, Jones RC, Karkanias J *et al.* The tabula sapiens: a multiple-organ, single-cell transcriptomic atlas of humans. *Science* 2022;**376**:eabl4896.
4. Guo D, Yang D, Zhang H, Song J, Wang P, Zhu Q *et al* DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*. 2025;**645**:633–8. https://doi.org/10.1038/s41586-025-09422-z
5. Zhao WX, Zhou K, Li J, Tang T, Dong Z, Hou Y *et al* A Survey of Large Language Models. *Frontiers of Computer Science*. 2026;**20**:2012627. https://doi.org/10.1007/s11704-026-60308-3
6. Zhou C, Li Q, Li C *et al.* A comprehensive survey on pretrained foundation models: a history from bert to chatgpt. *Int J Mach Learn Cybern* 2025;**16**:9851–15. https://doi.org/10.1007/s13042-024-02443-6
7. Touvron H, Lavril T, Izacard G *et al.* Llama: open and efficient foundation language models. *arXiv* 2023;**2302**:13971. https://doi.org/10.48550/arXiv.2302.13971
8. Cui H, Wang C, Maan H *et al.* scGPT: toward building a foundation model for single-cell multi-omics using generative AI. *Nat Methods* 2024;**21**:1470–80. https://doi.org/10.1038/s41592-024-02201-0
9. Theodoris CV, Xiao L, Chopra A *et al.* Transfer learning enables predictions in network biology. *Nature* 2023;**618**:616–24. https://doi.org/10.1038/s41586-023-06139-9
10. Hao M, Gong J, Zeng X *et al.* Large-scale foundation model on single-cell transcriptomics. *Nat Methods* 2024;**21**:1481–91. https://doi.org/10.1038/s41592-024-02305-7
11. Yang X, Liu G, Feng G *et al.* GeneCompass: deciphering universal gene regulatory mechanisms with a knowledge-informed cross-species foundation model. *Cell Res* 2024;**34**:830–45. https://doi.org/10.1038/s41422-024-01034-y
12. Wu J, Ye Q, Wang Y *et al.* Biology-driven insights into the power of single-cell foundation models. *Genome Biol* 2025;**26**:334. https://doi.org/10.1186/s13059-025-03781-6
13. Brodin P, Davis MM. Human immune system variation. *Nat Rev Immunol* 2017;**17**:21–9. https://doi.org/10.1038/nri.2016.125
14. Wu Q, Ding J, He R, Hui L, Liu J, Li Y. Exploring phenotype-related single-cells through attention-enhanced representation learning. *Genome Medicine*. 2026;**18**:21. https://doi.org/10.1186/s13073-026-01598-x
15. Martorell-Marugán J, López-Domínguez R, Villatoro-García JA *et al.* Explainable deep neural networks for predicting sample phenotypes from single-cell transcriptomics. *Brief Bioinform* 2025;**26**:bbae673. https://doi.org/10.1093/bib/bbae673
16. He B, Thomson M, Subramaniam M *et al.* CloudPred: predicting patient phenotypes from single-cell RNA-seq. *Biocomputing 2022 (WORLD SCIENTIFIC, Kohala Coast, Hawaii, USA.* Edited by: Russ B Altman, A Keith Dunke, Lawrence Hunter, Marylyn D Ritchie, Tiffany Murray, and Teri E Klein) 2022; pp. 337–48. https://doi.org/10.1142/9789811250477_0031
17. Zaheer M, Kottur S, Ravanbhakhsh S *et al.* Deep Sets. *Proceedings of the 31st International Conference on Neural Information Processing Systems* (Long Beach, California, USA: Curran Associates Inc.) 2017;p. 3394–404.
18. Molinari C, Marisi G, Passardi A *et al.* Heterogeneity in colorectal cancer: a challenge for personalized medicine? *Int J Mol Sci* 2018;**19**:3733. https://doi.org/10.3390/ijms19123733
19. Ahmad T, Lund LH, Rao P *et al.* Machine learning methods improve prognostication, identify clinically distinct phenotypes, and detect heterogeneity in response to therapy in a large cohort of heart failure patients. *J Am Heart Assoc* 2018;**7**:e008081. https://doi.org/10.1161/JAHA.117.008081
20. Chen L, Manautou JE, Rasmussen TP *et al.* Development of precision medicine approaches based on inter-individual variability of BCRP/ABCG2. *Acta Pharm Sin B* 2019;**9**:659–74. https://doi.org/10.1016/j.apsb.2019.01.007
21. Rivenbark AG, O'Connor SM, Coleman WB. Molecular and cellular heterogeneity in breast cancer: challenges for personalized medicine. *Am J Pathol* 2013;**183**:1113–24. https://doi.org/10.1016/j.ajpath.2013.08.002
22. Ren X, Wen W, Fan X *et al.* COVID-19 immune features revealed by a large-scale single-cell transcriptome atlas. *Cell* 2021;**184**:1895–913.e19. https://doi.org/10.1016/j.cell.2021.01.053
23. Popescu M-C, Balas VE, Perescu-Popescu L *et al.* Multilayer perceptron and neural networks. *WSEAS Trans Circuit Syst* 2009;**8**:579–88.
24. He K, Chen X, Xie S *et al.* Masked autoencoders are scalable vision learners. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 2022;p.15979–88. https://doi.org/10.1109/CVPR52688.2022.01553
25. Dosovitskiy A, Beyer L, Kolesnikov A *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv* 2020;**2010**:11929. https://doi.org/10.48550/arXiv.2010.11929
26. Ahern DJ, Ai Z, Ainsworth M *et al.* A blood atlas of COVID-19 defines hallmarks of disease severity and specificity. *Cell* 2022;**185**:916–38.e58. https://doi.org/10.1016/j.cell.2022.01.012
27. Stephenson E, Reynolds G, Botting RA *et al.* Single-cell multi-omics analysis of the immune response in COVID-19. *Nat Med* 2021;**27**:904–16. https://doi.org/10.1038/s41591-021-01329-2
28. Yoshida M, Worlock KB, Huang N *et al.* Local and systemic responses to SARS-CoV-2 infection in children and adults. *Nature* 2022;**602**:321–7. https://doi.org/10.1038/s41586-021-04345-x
29. van der Wijst MGP, Vazquez SE, Hartoularos GC *et al.* Type I interferon autoantibodies are associated with systemic immune alterations in patients with COVID-19. *Sci Transl Med* 2021;**13**:eabh2624. https://doi.org/10.1126/scitranslmed.abh2624
30. Rabadam G, Wibrand C, Flynn E *et al.* Coordinated immune dysregulation in juvenile dermatomyositis revealed by single-cell genomics. *JCI Insight* 2024;**9**:e176963. https://doi.org/10.1172/jci.insight.176963
31. Yang Y, Chen X, Pan J *et al.* Pan-cancer single-cell dissection reveals phenotypically distinct B cell subtypes. *Cell* 2024;**187**:4790–811.e22. https://doi.org/10.1016/j.cell.2024.06.038
32. Zhang J-Y, Wang X-M, Xing X *et al.* Single-cell landscape of immunological responses in patients with COVID-19. *Nat Immunol* 2020;**21**:1107–18. https://doi.org/10.1038/s41590-020-0762-x
33. Carr EJ, Dooley J, Garcia-Perez JE *et al.* The cellular composition of the human immune system is shaped by age and cohabitation. *Nat Immunol* 2016;**17**:461–8. https://doi.org/10.1038/ni.3371
34. Noh JY, Jeong HW, Kim JH *et al.* T cell-oriented strategies for controlling the COVID-19 pandemic. *Nat Rev Immunol* 2021;**21**:687–8. https://doi.org/10.1038/s41577-021-00625-9
35. Wherry EJ, Barouch DH. T cell immunity to COVID-19 vaccines. *Science* 2022;**377**:821–2. https://doi.org/10.1126/science.add2897
36. Wang Q, Zhu H, Hu Y *et al.* CellMemory: hierarchical interpretation of out-of-distribution cells using bottlenecked transformer. *Genome Biol* 2025;**26**:178. https://doi.org/10.1186/s13059-025-03638-y

37. Lotfollahi M, Yuhan H, Theis FJ *et al.* The future of rapid and automated single-cell data analysis using reference mapping. *Cell* 2024;**187**:2343–58. https://doi.org/10.1016/j.cell.2024.03.009
38. Dann E, Cujba A-M, Oliver AJ *et al.* Precise identification of cell states altered in disease using healthy single-cell references. *Nat Genet* 2023;**55**:1998–2008. https://doi.org/10.1038/s41588-023-01523-7
39. Stringer E, Bohnsack J, Bowyer SL *et al.* Treatment approaches to juvenile dermatomyositis (JDM) across North America: the childhood arthritis and rheumatology research alliance (CARRA) JDM treatment survey. *J Rheumatol* 2010;**37**:1953–61. https://doi.org/10.3899/jrheum.090953
40. Khojah A, Pachman LM, Bukhari A *et al.* Decreased peripheral blood natural killer cell count in untreated juvenile dermatomyositis is associated with muscle weakness. *Int J Mol Sci* 2024;**25**:7126. https://doi.org/10.3390/ijms25137126
41. Nagy N, Kuipers HF, Marshall PL *et al.* Hyaluronan in immune dysregulation and autoimmune diseases. *Matrix Biol* 2019;**78-79**:292–313. https://doi.org/10.1016/j.matbio.2018.03.022
42. Poole AR, Witter J, Roberts N *et al.* Inflammation and cartilage metabolism in rheumatoid arthritis. Studies of the blood markers hyaluronic acid, orosomucoid, and keratan sulfate. *Arthritis Rheum* 1990;**33**:790–9. https://doi.org/10.1002/art.1780330605
43. Maurice DH, Ke H, Ahmad F *et al.* Advances in targeting cyclic nucleotide phosphodiesterases. *Nat Rev Drug Discov* 2014;**13**:290–314. https://doi.org/10.1038/nrd4228
44. Jakubek YA, Chang K, Sivakumar S *et al.* Large-scale analysis of acquired chromosomal alterations in non-tumor samples from patients with cancer. *Nat Biotechnol* 2020;**38**:90–6. https://doi.org/10.1038/s41587-019-0297-6
45. Aran D, Camarda R, Odegaard J *et al.* Comprehensive analysis of normal adjacent to tumor transcriptomes. *Nat Commun* 2017;**8**:1077. https://doi.org/10.1038/s41467-017-01027-z
46. Gadaleta E, Fourgoux P, Pirró S *et al.* Characterization of four subtypes in morphologically normal tissue excised proximal and distal to breast cancer. *NPJ Breast Cancer* 2020;**6**:38. https://doi.org/10.1038/s41523-020-00182-9
47. Gorlova OY, Gorlov IP, Ripley RT *et al.* Gene expression in tumor and adjacent normal tissues in lung adenocarcinoma subtypes. *BMC Cancer* 2025;**25**:1169. https://doi.org/10.1186/s12885-025-14496-z
48. Zhu H, Lin Y, Lu D *et al.* Proteomics of adjacent-to-tumor samples uncovers clinically relevant biological events in hepatocellular carcinoma. *Natl Sci Rev* 2023;**10**:nwad167. https://doi.org/10.1093/nsr/nwad167
49. Wang H, Ma X. Learning deep features and topological structure of cells for clustering of scRNA-sequencing data. *Brief Bioinform* 2022;**23**:bbac068. https://doi.org/10.1093/bib/bbac068
50. Wang H, Ma X. Learning discriminative and structural samples for rare cell types with deep generative model. *Brief Bioinform* 2022;**23**:bbac317. https://doi.org/10.1093/bib/bbac317
51. Tan D, Yang C, Wang J *et al.* scAMAC: self-supervised clustering of scRNA-seq data based on adaptive multi-scale autoencoder. *Brief Bioinform* 2024;**25**:bbae068. https://doi.org/10.1093/bib/bbae068
52. Liao M, Liu Y, Yuan J *et al.* Single-cell landscape of bronchoalveolar immune cells in patients with COVID-19. *Nat Med* 2020;**26**:842–4. https://doi.org/10.1038/s41591-020-0901-9
53. Huang L, Shi Y, Gong B *et al.* Dynamic blood single-cell immune responses in patients with COVID-19. *Signal Transduct Target Ther* 2021;**6**:110. https://doi.org/10.1038/s41392-021-00526-2
54. Crook H, Raza S, Nowell J *et al.* Long covid—mechanisms, risk factors, and management. *BMJ* 2021;**374**:n1648. https://doi.org/10.1136/bmj.n1648
55. Raveendran A, Jayadevan R, Sashidharan S. Long COVID: an overview. *Diabetes Metab Syndr* 2021;**15**:869–75. https://doi.org/10.1016/j.dsx.2021.04.007
56. Banchereau R, Hong S, Cantarel B *et al.* Personalized immunomonitoring uncovers molecular networks that stratify lupus patients. *Cell* 2016;**165**:551–65. https://doi.org/10.1016/j.cell.2016.03.008
57. Rebouissou S, Amessou M, Couchy G *et al.* Frequent in-frame somatic deletions activate gp130 in inflammatory hepatocellular tumours. *Nature* 2009;**457**:200–4. https://doi.org/10.1038/nature07475
58. Stahl EA, Raychaudhuri S, Remmers EF *et al.* Genome-wide association study meta-analysis identifies seven new rheumatoid arthritis risk loci. *Nat Genet* 2010;**42**:508–14. https://doi.org/10.1038/ng.582
59. Bombardier C, Gladman DD, Urowitz MB *et al.* Derivation of the sledai. A disease activity index for lupus patients. *Arthritis Rheum* 1992;**35**:630–40. https://doi.org/10.1002/art.1780350606
60. Bode RK, Klein-Gitelman MS, Miller ML *et al.* Disease activity score for children with juvenile dermatomyositis: reliability and validity evidence. *Arthritis Rheum* 2003;**49**:7–15. https://doi.org/10.1002/art.10924
61. van der Heijde DM, van 't Hof MA, van Riel PL *et al.* Judging disease activity in clinical practice in rheumatoid arthritis: first step in the development of a disease activity score. *Ann Rheum Dis* 1990;**49**:916–20. https://doi.org/10.1136/ard.49.11.916
62. Weber CE, Kuo PC. The tumor microenvironment. *Surg Oncol* 2012;**21**:172–7. https://doi.org/10.1016/j.suronc.2011.09.001
63. Patil NS, Nabet BY, Müller S *et al.* Intratumoral plasma cells predict outcomes to PD-L1 blockade in non-small cell lung cancer. *Cancer Cell* 2022;**40**:289–300.e4. https://doi.org/10.1016/j.ccell.2022.02.002
64. Shalapour S, Font-Burgada J, Di Caro G *et al.* Immunosuppressive plasma cells impede T-cell-dependent immunogenic chemotherapy. *Nature* 2015;**521**:94–8. https://doi.org/10.1038/nature14395

Briefings in Bioinformatics, 2026, **27**, bbag286
https://doi.org/10.1093/bib/bbag286
Problem Solving Protocol