npj | systems biology and applications

Article

Published in partnership with the Systems Biology Institute

https://doi.org/10.1038/s41540-024-00484-9

A joint analysis of single cell transcriptomics and proteomics using transformer

Check for updates

Yuanyuan Chen[1], Xiaodan Fan[2], Chaowen Shi[3], Zhiyan Shi[1] & Chaojie Wang[1,4] ✉

CITE-seq provides a powerful method for simultaneously measuring RNA and protein expression at the single-cell level. The integrated analysis of RNA and protein expression in identical cells is crucial for revealing cellular heterogeneity. However, the high experimental costs associated with CITE-seq limit its widespread application. In this paper, we propose scTEL, a deep learning framework based on Transformer encoder layers, to establish a mapping from sequenced RNA expression to unobserved protein expression in the same cells. This computation-based approach significantly reduces the experimental costs of protein expression sequencing. We are now able to predict protein expression using single-cell RNA sequencing (scRNA-seq) data, which is well-established and available at a lower cost. Moreover, our scTEL model offers a unified framework for integrating multiple CITE-seq datasets, addressing the challenge posed by the partial overlap of protein panels across different datasets. Empirical validation on public CITE-seq datasets demonstrates scTEL significantly outperforms existing methods.

Over the past two decades, the rapid development of multi-omics sequencing technologies has profoundly transformed our understanding of cell biology. These technologies enable the sequencing of the genome, epigenome, transcriptome, proteome, and metabolome at the single-cell level, offering new insights into the interplay between intracellular and intercellular molecular mechanisms that govern development, physiology and pathogenesis[1]. Recent advancements in mono-omics research, particularly through single-cell RNA sequencing (scRNA-seq) methods, have already evolved to revolutionize our knowledge of cell types as well as their different functional cell states[2,3]. However, mono-omics alone is insufficient to fully unravel the molecular hierarchy from the genome to the phenome in individual cells. Multi-omics approaches at the single-cell level, which integrate multiple types of biological datasets, are essential and have emerged as a primary trend in cell biology research[4].

CITE-Seq (Cellular Indexing of Transcriptomes and Epitopes by Sequencing) is a cutting-edge method that simultaneously conducts mRNA sequencing and profiles surface proteins using available antibodies at the single-cell level[5]. This technique provides both proteomics and transcriptomics data for the same cell. Transcriptional information is helpful to understand cell physiology, but it cannot fully capture cellular functions without protein expression information. For example, RNA analysis is hard to explain post-transcriptional and post-translational modifications such as protein degradation, isoform detection, and glycosylation[6,7]. Indeed, proteins are fundamental to all aspects of cellular function. They play a critical role in shaping the cellular architecture and catalyzing biochemical reactions as enzymes[8,9]. Additionally, proteins are directly involved in crucial processes such as cellular signaling and intercellular interactions[10,11]. This highlights the importance of proteins in maintaining and regulating cellular activities. Nowadays, CITE-seq offers a powerful method for simultaneously profiling surface proteins and mRNA within single cells, thus providing a comprehensive view of cellular functions. Recently, CITE-Seq has facilitated several important discoveries in cell biology. Through profiling immune cells, Su et al. (2020)[12] identify a major shift between mild and moderate COVID-19 disease by analyzing CITE-Seq datasets. Additionally, Revelo et al. (2021)[13] found that a heterogeneous population of macrophages can prevent heart damage by conducting CITE-seq of cardiac immune cells. Wu et al. (2021)[14] utilized CITE-Seq to classify breast cancer cells based on their cellular composition and responses to treatments, providing a "comprehensive transcriptional atlas" that helps unravel the complex heterogeneity of breast cancer cells.

While CITE-seq offers a robust framework for multi-omics expression profiling, its widespread application still faces several challenges. The high experimental cost of generating CITE-seq data remains a significant barrier, especially for laboratories operating on limited budgets[15]. Moreover, the

[1]School of Mathematical Science, Jiangsu University, Zhenjiang, 212013 Jiangsu, China. [2]Department of Statistics, The Chinese University of Hong Kong, Shatin, N.T. Hong Kong, SAR, China. [3]School of Life Sciences, Jiangsu University, Zhenjiang, 212013 Jiangsu, China. [4]The Fourth Affiliated Hospital of Jiangsu University, Jiangsu University, Zhenjiang, 212013 Jiangsu, China. ✉e-mail: cjwang@ujs.edu.cn

limited availability of specific antibodies restricts the range of surface proteins that can be simultaneously measured. Issues such as antibody cross-reactivity and nonspecific binding can lead to spurious results, raising the risk of false-positive findings in CITE-seq experiments[16]. These technical artifacts complicate the interpretation of data and could skew biological conclusions. Furthermore, the correlation between RNA and surface protein expression levels is often weak due to various biological factors, such as post-transcriptional and post-translational modifications[17], adding further complexity to the analysis of CITE-seq data.

To address the limitations inherent in CITE-seq, several computation-based approaches have been proposed to infer the relationship between RNA and protein expression at the single-cell level. In fact, if protein expression can be accurately mapped from scRNA-seq data within the same cell, many of the challenges could be resolved, as the scRNA-seq technique is well-established and its costs have become more affordable. Currently, the commonly used workflows for integrating transcriptomic and proteomic data include Seurat[18] and totalVI[19]. Seurat is a popular R package for the analysis of single-cell data. It offers a comprehensive suite of tools for preprocessing, normalization, clustering, dimensionality reduction, and visualization, making it suitable for various data types, including CITE-seq. On the other hand, totalVI (Total Variational Inference) is specifically designed for the integrated analysis of scRNA-seq and protein expression data, such as those obtained from CITE-seq and REAP-seq (RNA Expression and Protein sequencing). TotalVI employs a unified probabilistic framework based on variational inference and Bayesian methods to model both RNA and protein measurements from single cells. By exploring the relationship between RNA and protein expression with reference datasets, Seurat and totalVI are capable of predicting the levels of surface proteins from a given scRNA-seq dataset.

Although Seurat and totalVI provide methods for integrating data from different sources or batches, they cannot fully correct for batch effects, particularly when consolidating multiple CITE-seq datasets with partially overlapping protein panels[20]. Moreover, as with any model-based approach, totalVI relies on certain assumptions about data distribution and structure, which may not always align perfectly with the actual data. Consequently, recent studies are turning to deep learning frameworks to jointly model scRNA-seq and protein expression data[21,22]. Lakkis et al. (2022)[23] proposed a versatile deep learning framework, sciPENN, which utilizes the architecture of recurrent neural networks (RNNs) to perform multiple tasks, including predicting protein expression for scRNA-seq, imputing protein expression for CITE-seq, and transferring cell type labels from CITE-seq to scRNA-seq. However, since the RNN block was primarily designed for sequential data, it may not be ideally suited for modeling expression matrix data. Furthermore, RNN models are known to suffer from several issues, such as gradient vanishing during the training process[24], and generally underperform compared to their revised version, long short-term memory (LSTM) models[25]. These limitations restrict the performance of the sciPENN model across various tasks.

In 2017, Vaswani et al. (2017)[26] proposed the groundbreaking Transformer architecture, which utilizes attention mechanisms instead of traditional CNN and RNN structures. In recent years, the Transformer architecture has become a cornerstone in deep learning, especially with the development of large language models (LLM)[27]. Inspired by the success of Transformer, this paper presents a novel framework named scTEL, which combines the Transformer Encoder layers with LSTM cells to jointly model transcriptomic and proteomic data in CITE-seq. Different from the RNN architecture in sciPENN, our scTEL model utilizes Transformer Encoder modules to extract embedding information from gene expression data. Through the attention mechanism, the Transformer Encoder can capture the underlying interrelationships among genes more effectively than simple linear layers. Additionally, scTEL combines a more interpretable LSTM architecture to enable a multi-task framework. Through empirical evaluations across multiple public datasets, we demonstrate that our scTEL model outperform existing approaches in protein expression prediction, cell type identification, and data integration. To the best of our knowledge, scTEL provides the most effective framework for analyzing CITE-seq data at the single-cell level.

This paper is organized as follows: Section “Methods” describes the datasets used, details the structure of scTEL, and outlines the training process. Section “Results” presents the performance of models in protein expression prediction, cell type identification, and data integration. Section “Discussion” further discusses the results.

Methods

Datasets and normalization

The CITE-seq technique allows for simultaneous sequencing of mRNA and protein expression at the single-cell level. A CITE-seq dataset comprises both an RNA expression matrix and a protein expression matrix for the same cells (refer to Fig. 1a). Due to the high experimental costs, published CITE-seq datasets are relatively scarce. To evaluate the performance of our model, this paper considers the following four publicly available datasets as benchmarks: the human Peripheral Blood Mononuclear Cells dataset (PBMC)[28], the Mucosa-Associated Lymphoid Tissue dataset (MALT)[23], the human Blood Monocyte and Dendritic cell CITE-seq dataset (Monocytes)[23], and the H1N1 influenza PBMCs dataset[29] (H1N1). These datasets were sequenced using different sequencing platforms, demonstrating the broad applicability of our method across various datasets. Sourced from reputable institutions, they ensure data integrity and reliable cell type labeling. Consequently, many related studies use them as benchmarks to evaluate model performance[23,30–32], as does this paper. Table 1 presents a brief summary of the four datasets analyzed in this study.

Before inputting the datasets into the model, we perform UMI (unique molecular identifier) normalization on the expression matrices for RNA genes and proteins, respectively[33]. Normalization is an essential preprocessing step in scRNA-seq data analyses. Its primary goal is to mitigate the impact of technical artifacts, such as experimental noise and sequencing depth, on molecular counts while preserving true biological variation[34]. In this paper, UMI normalization is performed using the Python package Scanpy[35].

Specifically, we begin by dividing the UMI counts by the total UMI counts in each cell, followed by multiplication with the median of the total UMI counts across all cells. Subsequently, we take the natural logarithm of these normalized UMI counts:

$$v_{ij} = \log\left(\frac{u_{ij}}{\sum_{j=1}^{g} u_{ij}} \cdot \text{median}\,(\mathbf{U}) + 1\right),$$

where $\mathbf{U} = \{u_{ij}\}_{n\times g}$ represents the original expression matrix with n cells and g genes, and v_{ij} denotes the log-normalized UMI counts for the j-th gene in the i-th cell.

Finally, we utilize z-score normalization to ensure that the mean expression level for each gene is 0 and the standard deviation is 1:

$$x_{ij} = \frac{v_{ij} - \mu_j}{\sigma_j},$$

where $\mu_j = \frac{1}{n}\sum_{i=1}^{n} v_{ij}$ and $\sigma_j = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(v_{ij} - \mu_j)^2}$ represent the sample mean and standard deviation of the log-normalized expression for each gene, respectively. The resulting matrix $\mathbf{X} = \{x_{ij}\}_{n\times g}$ becomes the gene expression matrix after UMI normalization. Similarly, let $\mathbf{Y} = \{y_{ij}\}_{n\times p}$ represent the protein expression matrix after UMI normalization following the same steps, where p is the number of proteins involved. Through UMI normalization, our goal is to mitigate the impact of technical noise and bias, thereby ensuring the comparability of expression data across different cells.

scTEL architecture

Overview. The primary goal of scTEL is to establish a mapping from the RNA expression matrix to protein expression at the single-cell level. This mapping enables effective prediction of unobserved protein expressions using scRNA-seq data, which is well-established and available at a low

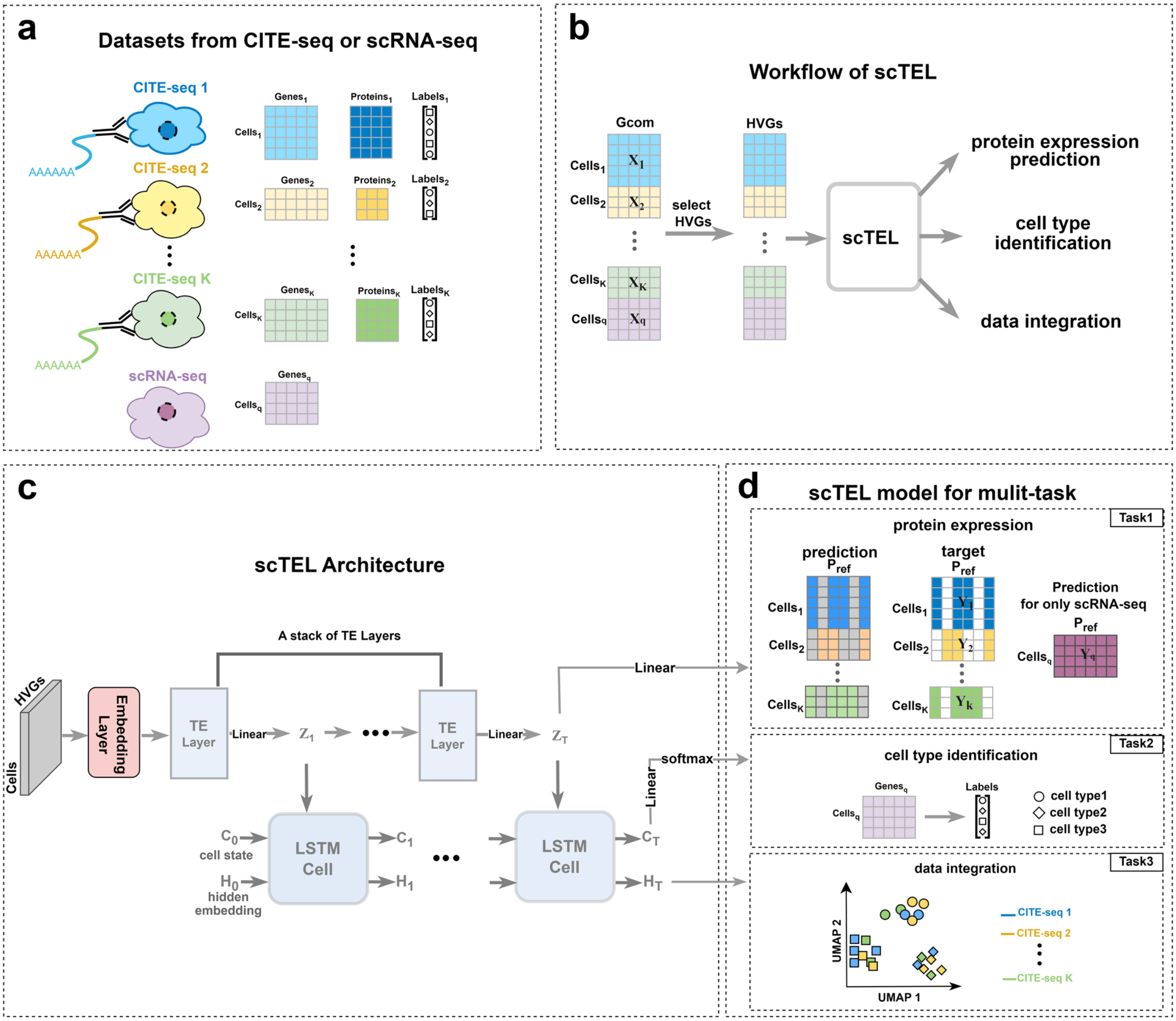

Fig. 1 | Overview of scTEL and its multiple tasks. a The CITE-seq dataset consists of an RNA expression matrix and a protein expression matrix for the same cells, along with the actual labels indicating the cell types for each individual cell. **b** Multiple RNA expression matrices after UMI normalization are merged together and filtered by HVGs. Then, the HVG expression matrix is fed into scTEL to accomplish three tasks. **c** The architecture of scTEL is composed of an embedding layer followed by a stack of TE layers integrated with LSTM cells. **d** The scTEL model is designed for three important tasks: protein expression prediction, cell type identification, and data integration.

Table 1 | Sample size and sequencing information of datasets used for evaluation

Dataset	Cells	Genes	Proteins	Sequencing platform
PBMC	161,764	20,729	224	Illumina NovaSeq 6000
MALT	8412	33,538	17	Illumina HiSeq 2500
H1N1	53,201	32,738	87	10x Genomics Chromium Controller
Monocytes	37,112	22,060	283	BD FACSAri II

cost. CITE-seq datasets serve as reference benchmarks for elucidating the relationship between RNA and protein expressions in the same cells.

Given multiple CITE-seq datasets after normalization, we integrate them by identifying common genes across all datasets. Let

$$G_{\mathrm{com}} = G_1 \cap G_2 \cap \cdots \cap G_K,$$

where G_k, $k = 1, 2, \cdots, K$, represents the set of genes in the k-th CITE-seq dataset, and G_{com} denotes the set of common genes across all K CITE-seq datasets. We then select the top 1,000 highly variable genes (HVGs) from these common genes as the input features. Through embedding layers, Transformer encoding (TE) layers, and LSTM cells, the scTEL modol outputs an integrated protein expression matrix that encompasses the union of protein sets from all datasets. Let

$$P_{\mathrm{ref}} = P_1 \cup P_2 \cup \cdots \cup P_K,$$

where P_k, $k = 1, 2, \cdots, K$, represents the set of proteins in the k-th CITE-seq dataset and P_{ref} denotes the union of all protein sets, which is called reference proteins.

Because the protein panels in different CITE-seq datasets may only partially overlap, merging the protein expression matrices will inevitably result in empty entries. These empty entries in the target matrix will not be taken into account in the calculation of the training loss. Figure 1b presents the overview workflow of the scTEL model briefly.

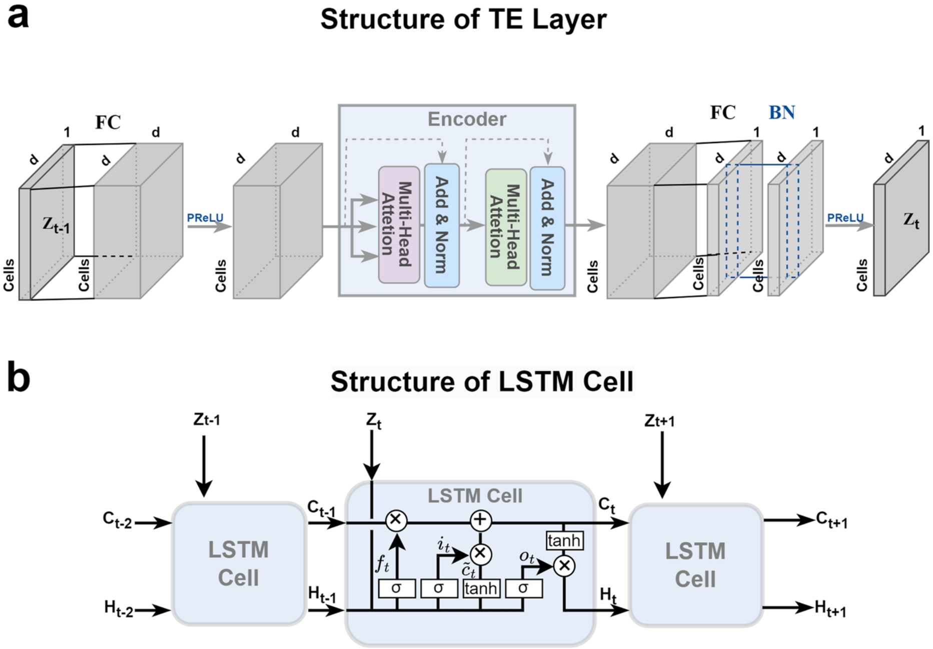

Fig. 2 | Detailed structures of TE layers and LSTM cells. a A TE layer consists of fully-connected embedding layers combined with an Encoder module. **b** A standard LSTM cell contains an input gate i_t, a forget gate f_t, and an output gate o_t.

Multi-tasks. Different from traditional methods tailored for specific tasks, scTEL provides a joint analysis framework suitable for multiple tasks. Through the combination of TE layers and LSTM cells, the scTEL model can accomplish protein expression prediction, cell type identification, and data integration simultaneously. These three tasks are inherently correlated. Effective data integration surely helps cell type identification, and further improves the accuracy of protein expression prediction. To capture this connection, the scTEL model generates multiple outputs based on shared embedding features to address these tasks.

Specifically, the RNA expression matrix, with highly variable genes as input data, is fed into the scTEL model and passes through the embedding layer. The structure of the embedding layer is shown in the Supplementary Fig. 1. This layer projects the high-dimensional data into a low-dimensional space to achieve dimension reduction. Subsequently, the embedded data pass through a stack of TE layers, generating a sequence of intermediate matrices $\mathbf{Z}_t$ with the embedding dimension. The feature extraction capabilities of the Transformer architecture have been widely validated across various fields[36]. In our approach, the TE layers are designed to extract embedded features from gene expression matrices. Using the attention mechanism, the TE layers effectively capture interaction information within the gene expression data, and generate intermediate representations (see Fig. 2a). This module offers substantial advantages over the simple linear layers used in sciPENN, which cannot fully capture the complex interrelationships among genes. Ultimately, the final matrix $\mathbf{Z}_T$ is linearly projected to predict the protein expression matrix (Task 1).

Additionally, the sequential intermediate matrices $\mathbf{Z}_t$ serve as inputs to LSTM cells, which are a variant of RNNs specialized in handling sequences with long-range dependencies (see Fig. 2b). LSTM cells are employed to capture and characterize the sequential information in the outputs of the TE layers. The multiple outputs of the LSTM cells align with our multi-task framework. Specifically, the LSTM cell state C_T is used to predict the cell type through linear layers and softmax functions (Task 2), while the hidden embedding H_T serves as the low-dimensional representation of the cells. By using linear projection and the Uniform Manifold Approximation and Projection (UMAP) method[37], we can visualize the latent space of the datasets to demonstrate the effectiveness of data integration across different CITE-seq datasets (Task 3). Figure 1c, d detail the architecture of the scTEL model and the corresponding tasks.

Structure of TE layers and LSTM cells. Inspired by classical Transformer architecture, our TE layer employs the attention mechanism to extract key latent features in the spatial structure of data. The attention mechanism is the core of Transformer and may be the most exciting innovation in deep learning in the past few years. The attention mechanism is inspired by the mechanism of human vision that our eyes often focus limited attention on the important local areas rather than the whole scope. Thus, it can save computational resources and capture the essential information quickly. The self-attention mechanism in Vaswani et al. (2017)[26] is defined as follows:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d}}\right)V$$

where $Q \in \mathbb{R}^{d\times d}$, $K \in \mathbb{R}^{d\times d}$, and $V \in \mathbb{R}^{d\times d}$ are the query, key and value matrices respectively, which are outputs of three different linear layers with the same input. The dot product of Q and K^T measures the similarity between query and key. Then, the attention is calculated by the weighted average of the corresponding value V.

Furthermore, we concatenate multiple self-attention together, called multi-head attention, to improve the performance. Each attention function is executed in parallel with the respective projected versions of the query, key, and value matrices. The multi-head attention can be expressed as follows:

$$\begin{aligned}\text{MultiHead}\,(Q, K, V) &= \text{Concat}(\text{head}_1, \text{head}_2, \cdots, \text{haed}_h)W^O,\\ \text{head}_i &= \text{Attention}\,(QW_i^Q, kW_i^K, VW_i^V),\end{aligned}$$

where $i = 1, \ldots, h$ and W_i^Q, W_i^K, W_i^V are weights of corresponding networks. The structure of the TE layer is shown in the Fig. 2a.

The sequential intermediate matrices $\mathbf{Z}_t$ serve as inputs to LSTM cells. The structure of the LSTM cell is shown in the Fig. 2b. Specifically, the feedforward steps in LSTM can be expressed as follows:

$$\begin{aligned}\boldsymbol{f}_t &= \sigma(\mathbf{W}_f\cdot[\mathbf{H}_{t-1}, \mathbf{Z}_t] + \boldsymbol{b}_f),\\ \boldsymbol{i}_t &= \sigma(\mathbf{W}_i\cdot[\mathbf{H}_{t-1}, \mathbf{Z}_t] + \boldsymbol{b}_i),\\ \boldsymbol{o}_t &= \sigma(\mathbf{W}_o\cdot[\mathbf{H}_{t-1}, \mathbf{Z}_t] + \boldsymbol{b}_o),\\ \mathbf{C}_t &= \boldsymbol{f}_t \otimes \mathbf{C}_{t-1} + \boldsymbol{i}_t \otimes \tanh\,(\mathbf{W}_c\cdot[\mathbf{H}_{t-1}, \mathbf{Z}_t] + \mathbf{b}_c),\\ \mathbf{H}_t &= \boldsymbol{o}_t \otimes \tanh\,(\mathbf{C}_t),\end{aligned}$$

where $\mathbf{C}_t$ represents the cell state and $\mathbf{H}_t$ represents the hidden embedding. Here, $\mathbf{W}_f$, $\mathbf{W}_i$, $\mathbf{W}_o$, $\mathbf{W}_c$, and $\boldsymbol{b}_f$, $\boldsymbol{b}_i$, $\boldsymbol{b}_o$, $\boldsymbol{b}_c$, are weight matrices and bias vectors for the corresponding connection, $\sigma(\cdot)$ and $\tanh(\cdot)$ represent the sigmoid function and the tanh function, $[\cdot\,,\cdot]$ denotes the concat operation which merges the two vectors together, and $\otimes$ denotes the Hadamard product of two vectors.

Loss function

In deep learning, the choice of loss functions traditionally depends on the nature of the task. For regression problems, where the targets are continuous variables, mean squared error (MSE) is a commonly used loss function. For classification problems involving discrete target variables, cross-entropy loss is the standard choice. Our scTEL model encompasses multiple tasks, including protein expression prediction as a regression problem and cell type identification as a classification problem. Each task requires a suitable loss function to measure the discrepancy between the model's predictions and the actual target values.

In Task 1, the scTEL model provides predictions for the expression matrix of reference proteins. Since the expression values are continuous, we consider using the MSE to measure the distance between the predicted expression matrix $\hat{\mathbf{Y}} = \{\hat{y}_{ij}\}$ and the actual expression matrix $\mathbf{Y} = \{y_{ij}\}$. Note that there are some empty entries in the true expression matrix due to the partial overlap of protein panels across different datasets. These empty entries are excluded from the loss calculation. Let $\mathbf{B} = \{b_{ij}\}$ be the indicator matrix with the same dimensions of $\mathbf{Y}$, where $b_{ij} = 1$ if the expression value of j-th protein in the i-th cell is sequenced in CITE-seq datasets; otherwise $b_{ij} = 0$. Therefore, the MSE loss is defined as:

$$L_{\mathrm{MSE}}(\mathbf{Y}, \hat{\mathbf{Y}}) = \frac{1}{n_{\mathrm{train}} \cdot p_{\mathrm{ref}}} \sum_{i=1}^{n_{\mathrm{train}}} \sum_{j=1}^{p_{\mathrm{ref}}} b_{ij}(y_{ij} - \hat{y}_{ij})^2,$$

where n_{train} denotes the number of cells in the training set and p_{ref} denotes the number of reference proteins. Here, n_{train} and p_{ref} are the dimensions of the expression matrix of reference proteins $\mathbf{Y}$.

Besides assessing the estimation error of the protein expression matrix, we also evaluate the uncertainty of the estimation by using the quantile loss function[38]. Let $Q = \{q_1, q_2, \cdots, q_M\}$ be the set of quantiles we wish to estimate. The final output $\boldsymbol{Z}_T$ is projected through linear layers to independently generate the m-th quantile estimates $\mathbf{Y}^{(m)} = \{y_{ij}^{(m)}\}$, for $m = 1, 2, \cdots, M$. The quantile loss function is defined as:

$$L_{\mathrm{quantile}} = \frac{1}{M \cdot n_{\mathrm{train}} \cdot p_{\mathrm{ref}}} \sum_{m=1}^{M} \sum_{i=1}^{n_{\mathrm{train}}} \sum_{j=1}^{p_{\mathrm{ref}}} b_{ij} \cdot l_q(y_{ij}, y_{ij}^{(m)}),$$

where $l_q(y_{ij}, y_{ij}^{(m)}) = [(1 - q_m)I(y_{ij}^{(m)} > y_{ij}) + q_m I(y_{ij}^{(m)} < y_{ij})] \cdot |y_{ij}^{(m)} - y_{ij}|$ denotes the entrywise quantile loss. Here, $I(\cdot)$ represents the indicator function, returning 1 if the inequality is true and 0 otherwise. This loss function evaluates the uncertainty of estimation and provides an overall view of the model's performance.

In Task 2, which is a classification task, the scTEL model utilizes a linear layer coupled with a softmax function as the activation function on the final cell state $\boldsymbol{C}_T$ to output probability predictions for cell types, assuming the number of cell types is predefined. Each cell is then identified as the cell type with the highest prediction probability. For this multiple classification task, the categorical cross-entropy (CE) loss function is a common choice to measure the distance between the predicted and true distributions[39]. Let $\boldsymbol{S} = (s_1, s_2, \cdots, s_{n_{\mathrm{train}}})$ represent the true cell type labels for the cells in the training set. The CE loss function can be defined as:

$$L_{\mathrm{CE}} = -\frac{1}{n_{\mathrm{train}}} \sum_{i=1}^{n_{\mathrm{train}}} \log(P_i(s_i)),$$

where $P_i(s_i)$ denotes the predicted probability that the i-th cell belongs to cell type s_i. This loss function focuses on maximizing the probability assigned to the correct class, thereby enhancing the model's accuracy in cell type identification.

Finally, we aggregate the three loss functions as the total loss, which serves as the objective in the training process that needs to be minimized:

$$L_{\mathrm{total}} = L_{\mathrm{MSE}} + L_{\mathrm{quantile}} + L_{\mathrm{CE}}.$$

By calculating the gradient of the loss function with respect to the model's parameters, the gradient descent algorithm can update the parameters in a direction that reduces the loss. The training loss is expected to converge to a lower level after sufficient epochs of iteration.

By minimizing the total loss, the scTEL model effectively captures the underlying connections between different tasks. It balances the objectives of each task and ensures that the trained model achieves optimal performance from a holistic perspective. For detailed information on the training process and model parameters, see the Supplementary Information.

Results

In this section, we design three scenarios using four public CITE-seq datasets, including PBMC, MALT, H1N1, and Monocytes, to evaluate the performance of the scTEL model. We compare scTEL with traditional methods such as Seurat and totalVI, as well as the deep learning framework sciPENN. Our results demonstrate that the scTEL model significantly outperforms these existing approaches across all tasks.

Data integration and low-dimensional representation

In single-cell analysis, each cell is characterized by high-dimensional gene sequences. However, technical noise and samples from different sources can introduce batch effects, which may confound the biological interpretation of the data. Therefore, integrating multiple datasets and removing batch effects are essential steps for further analysis.

UMAP is a widely used method for visualizing batch effects by projecting the high-dimensional expression data into a low-dimensional space, typically 2D or 3D[40]. After dimension reduction, each cell is plotted in the UMAP space and can be colored according to its batch label. This visual representation provides a clear indication of whether cells from different batches cluster separately (indicating batch effects) or intermingle (indicating successful mitigation of batch effects). Here, we explore three cases to evaluate the effectiveness of our scTEL model in data integration and batch effect correction by using UMAP visualization.

Case I: We divided the Monocytes dataset, which comprises 8 samples from 4 participants, into two subsets–Monocyte 1 and Monocyte 2, with each subset containing 4 samples from 2 participants. Since both subsets are sourced from the same laboratory and platform, the batch effect between the two subsets should be minimal. Figure 3a illustrates that most methods can integrate the two subsets well, with scTEL and sciPENN performing best by achieving complete data mixing. In contrast, totalVI shows some bias, and Seurat does not mix the subsets effectively.

Case II: We consider integrating the PBMC and H1N1 datasets. Although sourced from different laboratories, the cell types in these two datasets are closely related. Specifically, the H1N1 dataset comprises human peripheral blood mononuclear cells infected with H1N1 influenza. The difference between the PBMC and H1N1 datasets is more pronounced than that in Case I. In this scenario, Fig. 3b demonstrates that scTEL significantly outperforms other methods by achieving completely mixed plots. In contrast, totalVI and Seurat fail to achieve sufficient mixing, resulting in the H1N1 data clustering separately.

Case III: We consider integrating the PBMC and MALT datasets, which are completely distinct with minimal featuring similarity. This scenario represents the most challenging integration task. Figure 3c shows that only scTEL is able to mix the datasets effectively, while in other methods, the MALT cells cluster separately, indicating poor integration performance. This highlights the superior ability of scTEL to handle and integrate highly disparate datasets.

To further assess the performance of each batch effect correction method and demonstrate that data integration avoids over-integration, we

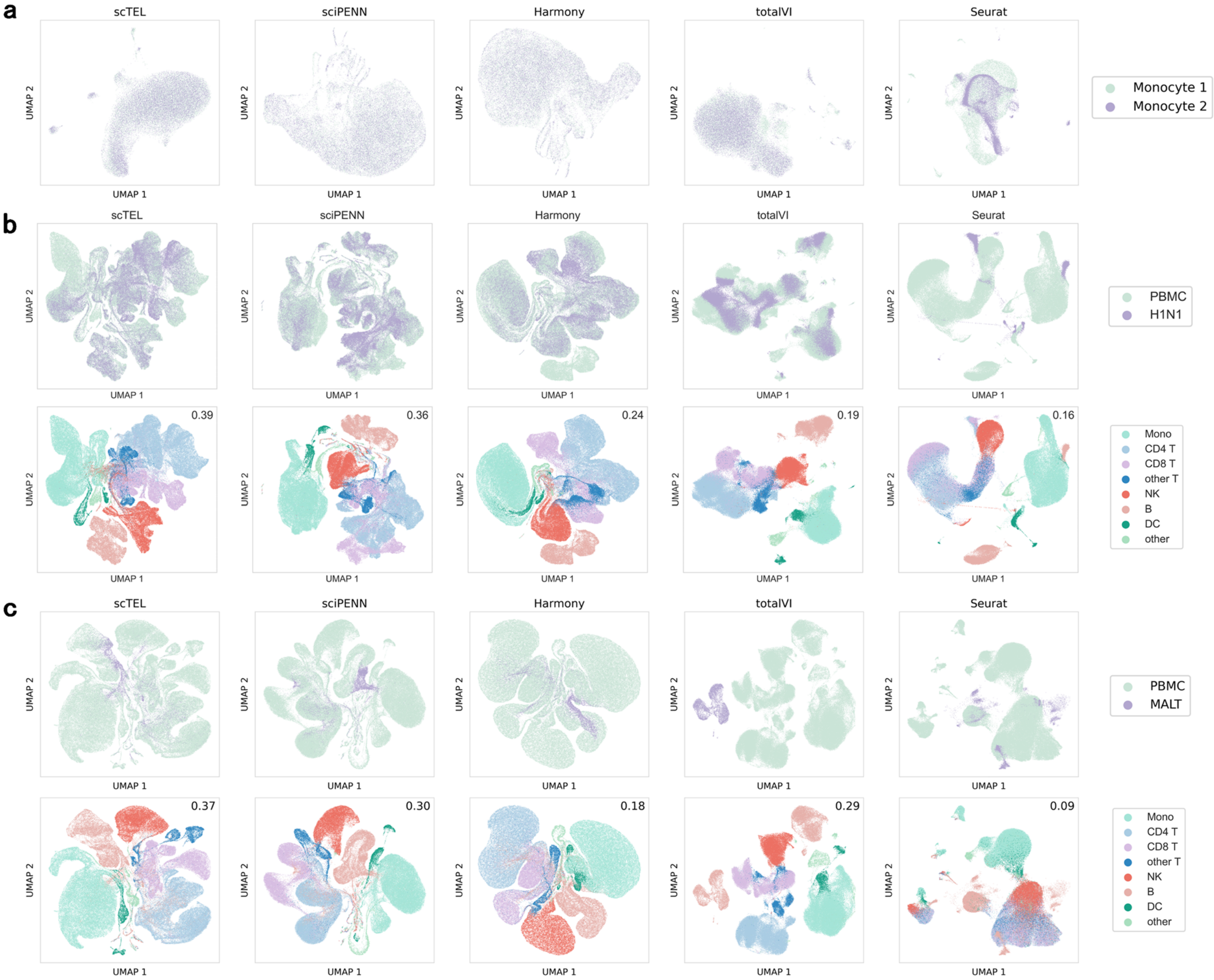

Fig. 3 | UMAP visualizations colored by datasets and cell annotations. It demonstrates the effectiveness of data integration across three scenarios for the five compared models. The Silhouette score, shown in the top-right corner of each plot, provides a quantitative measure of clustering performance for the different models. **a** Case I, **b** Case II, **c** Case III.

colored the true cell type labels in the PBMC and MALT datasets in Fig. 3b, c (note that true cell type labels are not available for the Monocytes dataset). Our results show that the two datasets are well-mixed in the scTEL model while maintaining clear and distinct clustering. Importantly, different clusters remain well-separated, indicating that the integration of the PBMC and MALT datasets avoids over-integration.

Furthermore, we employed the Silhouette score to quantitatively evaluate the clustering quality[41]. The results reveal that our scTEL model achieves the highest Silhouette scores across all cases, highlighting its superior performance.

From these cases, it is evident that as the difficulty of data integration increases, the scTEL model exhibits significant advantages over other methods. This demonstrates that scTEL can effectively handle data integration and mitigate batch effects, especially for highly disparate datasets.

Protein expression prediction

The primary goal of scTEL is to establish a mapping from RNA expression to protein expression. With a trained model, scTEL can impute the unsequenced protein expression in CITE-seq datasets, or predict a complete protein expression with scRNA-seq data (refer to Task 1 in Fig. 1d).

To evaluate the performance of models, both the root means square error (RMSE) and Pearson correlation are used to measure the distance between predicted and target expression levels for individual proteins within CITE-seq data[23]:

$$\begin{aligned}\mathrm{RMSE}_j &= \sqrt{\frac{1}{\sum_{i=1}^{n_{\mathrm{train}}} b_{ij}}\sum_{i=1}^{n_{\mathrm{train}}} b_{ij}(y_{ij}-\hat{y}_{ij})^2},\\ \rho_j &= \frac{\sum_{i=1}^{n_{\mathrm{train}}} b_{ij}(y_{ij}-\mu_j)(\hat{y}_{ij}-\hat{\mu}_j)}{\sqrt{\sum_{i=1}^{n_{\mathrm{train}}} b_{ij}(y_{ij}-\mu_j)^2}\sqrt{\sum_{i=1}^{n_{\mathrm{train}}} b_{ij}(\hat{y}_{ij}-\hat{\mu}_j)^2}},\end{aligned}$$

where RMSE_j and ρ_j denote the RMSE and Pearson correlation between the predicted and target protein expression levels for the j-th protein, respectively. Here, $\mu_j=\frac{1}{\sum_{i=1}^{n_{\mathrm{train}}} b_{ij}} b_{ij}y_{ij}$ and $\hat{\mu}_j=\frac{1}{\sum_{i=1}^{n_{\mathrm{train}}} b_{ij}} b_{ij}\hat{y}_{ij}$ are sample mean of target and predicted expression levels.

Figure 4a illustrates the box plot of the RMSE for each protein across four models in three distinct settings as outlined in Section "Data Integration and Low-dimensional Representation". Across all three settings, our scTEL model consistently exhibits the lowest average RMSE of all proteins. Especially in Case III, where PBMC and MALT datasets differ greatly, our scTEL model significantly outperforms other models.

Figure 4b visualizes the expression levels of 7 overlapping marker proteins in Monocytes, including CD14, CD74, CD36, CD62L, CD16, HLA-DR, and CD86. These proteins play critical roles in the immune system, participating in pathogen recognition, regulation of immune responses, and various cell-cell interaction functions. In immunology,

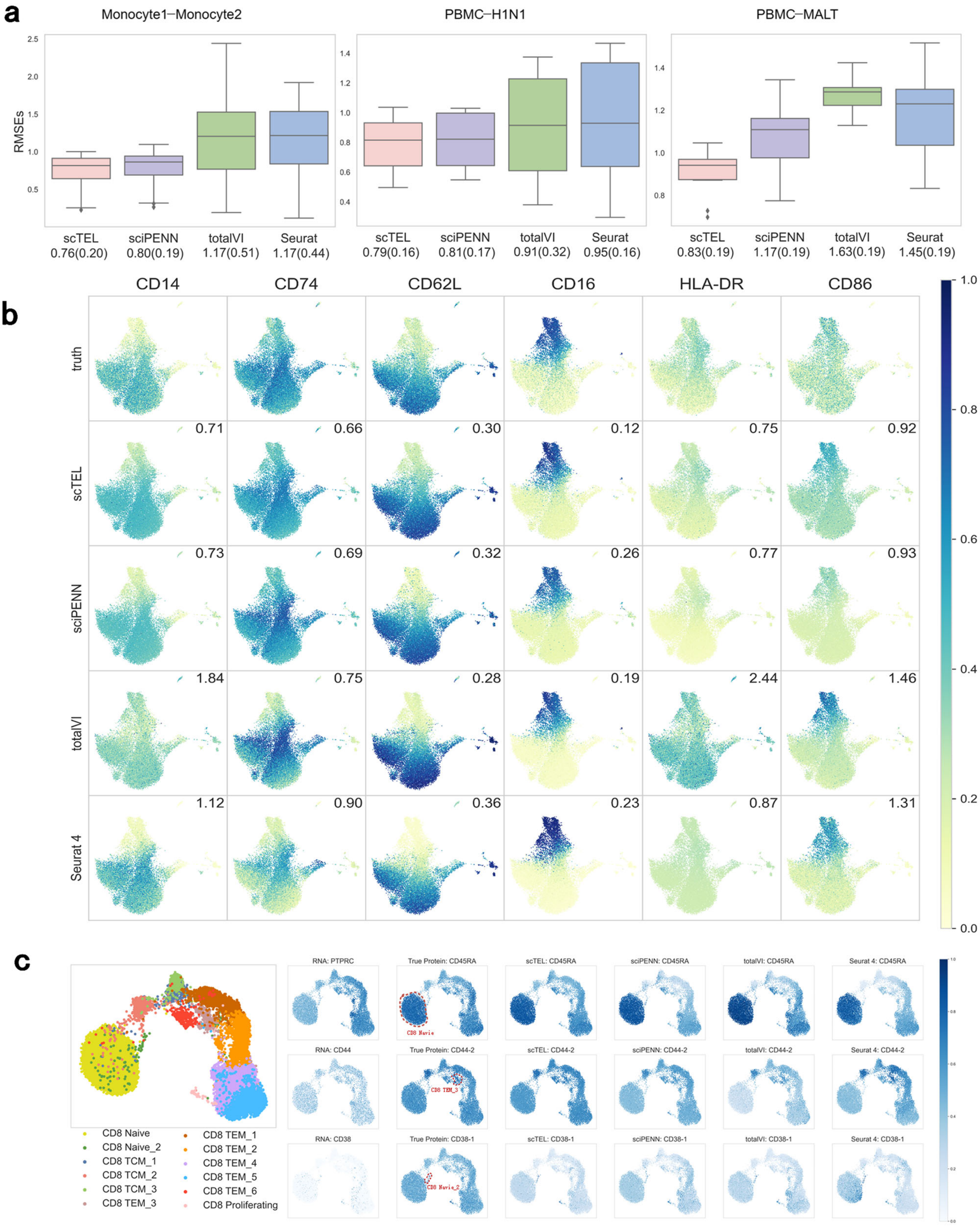

Fig. 4 | The RMSE performance of models for protein expression prediction. **a** Box plots of the RMSE for each protein across four models in three scenarios. The values at the bottom of the plot represent the means and standard errors of all proteins in each setting. **b** True and predicted expression levels of 6 overlapping marker proteins in Monocytes. The value indicated in the top-right corner of the plot corresponds to the RMSE for the specific protein in each setting. **c** The UMAP plot on the left illustrates the true labels of CD8 cell subpopulations in PBMC, while the UMAP plot on the right focuses on three specific subpopulations of CD8 cells: Naive, TEM3, and Naive2. By visualizing the expression patterns of marker proteins (CD45RA, CD44-2, and CD38-1) alongside their corresponding encoding RNA genes (PTPRC, CD44, and CD38), we demonstrate that protein expression predicted by scTEL enables the identification of specific cell subpopulations with high levels of particular protein expression.

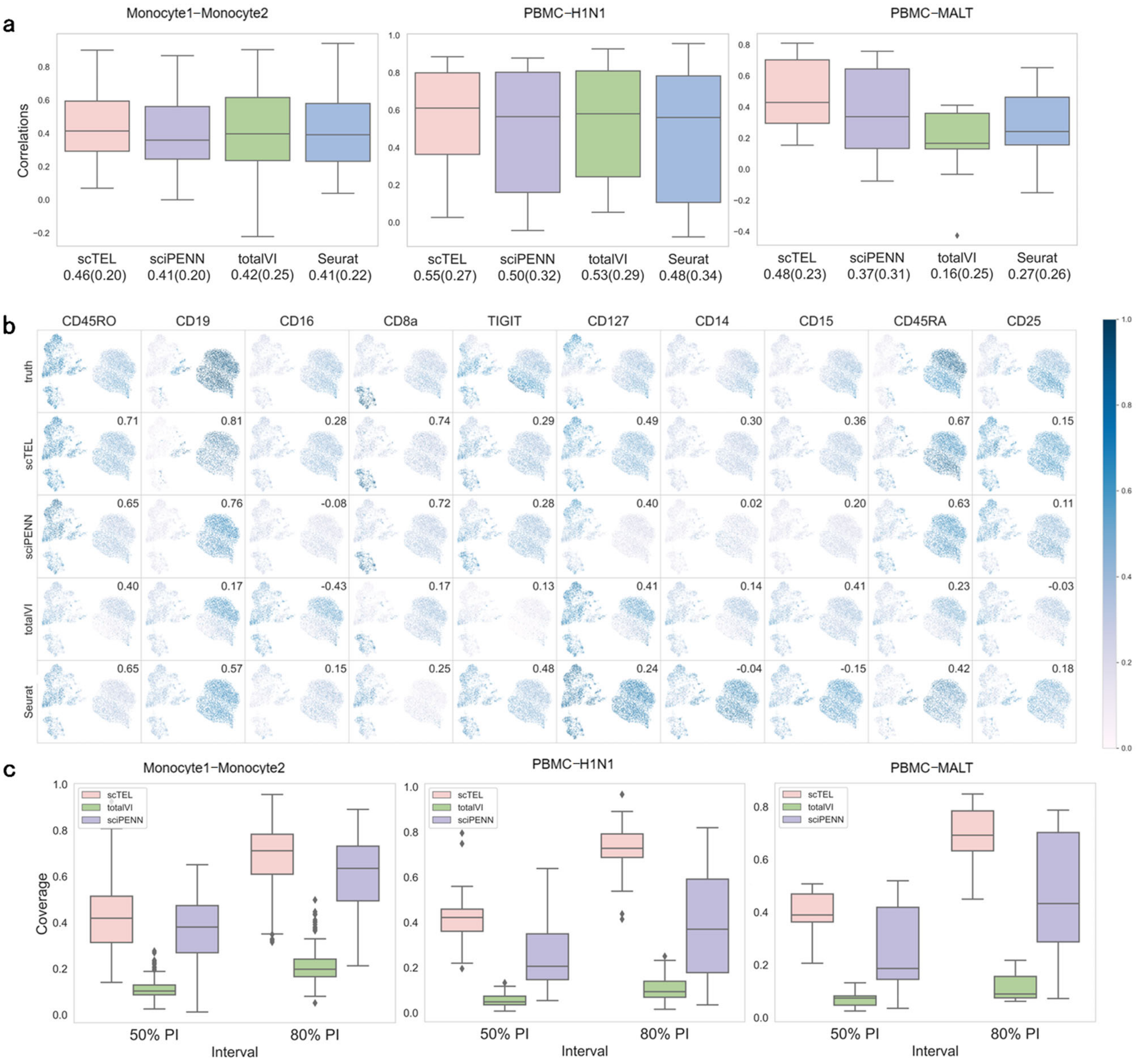

Fig. 5 | The Pearson correlation performance of models for protein expression prediction. a Box plots of the Pearson correlation for each protein across four models in three scenarios. The values at the bottom of the plot represent the means and standard errors of all proteins in each setting. **b** True and predicted expression levels of 10 marker proteins that are common to both PBMC and MALT datasets. The value indicated in the top-right corner of the plot corresponds to the correlation for the specific protein in each setting. **c** Box plots of coverage probabilities for proteins under 50% PI and 80% PI for the three models.

monocytes are classified into three subtypes–classical, non-classical, and intermediate–based on expression patterns of these proteins in immune cells[42,43]. Compared with other models, scTEL provides the closest predictions to the true values. This demonstrates that our scTEL model performs effectively across different subtypes of monocytes.

Figure 4c shows the gene expression alongside the corresponding protein expressions in PBMC. Analyzing the expression of coding RNA genes alone is insufficient to distinguish cell subpopulations effectively. However, incorporating protein expression data helps address this limitation and provides a more comprehensive understanding. As an example, for the three subpopulations of CD8 cells: Naive, TEM3, and Naive2. Using UMAP, we visualize the expression of marker proteins (CD45RA, CD44-2, and CD38-1, respectively) and their corresponding encoding RNA genes (PTPRC, CD44, and CD38). The results demonstrate that protein expression predicted by scTEL allows researchers to identify the specific cell subpopulations where particular proteins are highly expressed.

Similarly, Fig. 5 presents the Pearson correlation between predicted and target expression values, with higher correlations indicating better model performance. Figure 5(a) illustrates the box plot of the correlation for each protein across four models. In all three settings, our scTEL model consistently exhibits the highest average Pearson correlation of all proteins. Figure 5b visualizes the expression levels of 10 marker proteins that are common to both PBMC and MALT datasets. Compared with other models, scTEL provides the highest correlation with the true values for all proteins.

To consider the uncertainty of model predictions, we calculate the coverage probability, which indicates the proportion of true values that fall within the prediction intervals (PI), for individual proteins. A higher coverage probability reflects greater robustness of the models. As Seurat does not quantify protein expression prediction uncertainty, our comparison is limited to sciPENN and totalVI. Figure 5c depicts the box plot of coverage probabilities for proteins under 50% PI and 80% PI for the three models. Our scTEL model consistently outperforms the others in all scenarios. Table 2 provides a summary of the median coverage probability in each setting. As the task difficulty increases (from Case I to Case III), scTEL maintains robust performance, with the median coverage probability remaining stable across all settings. In contrast, sciPENN and totalVI experience a significant decline in performance, particularly in the more challenging Cases II and III.

Table 2 | Median of coverage probabilities under each setting

Cases	PI range	Median of coverage(%)		
		scTEL	sciPENN	totalVI
Monocyte1-Monocyte2	50% PI	**42.6**	38.1	10.3
	80% PI	**68.4**	63.5	19.8
PBMC-H1N1	50% PI	**42.0**	20.6	4.9
	80% PI	**72.8**	37.1	9.3
PBMC-MALT	50% PI	**39.6**	18.6	7.4
	80% PI	**69.0**	43.3	8.9

Cell type identification

Cell type identification and annotation constitute another critical aspect of single-cell analysis. Traditional scRNA-seq data analysis typically employ dimension reduction and clustering techniques on RNA expression matrices, followed by the identification of cell types using marker genes. However, certain cell types may not be easily distinguished based solely on RNA expression, particularly for some rare subtypes. Accurately distinguishing these cell subtypes is essential for capturing cellular heterogeneity and understanding the diverse functions within biological systems[44,45].

Our scTEL offers a powerful tool for identifying cell types, leveraging both RNA and protein expression data to enhance classification accuracy. In this section, we assess the performance of our model in cell type identification using PBMC datasets. TotalVI is excluded from the model comparison, as it is not designed for cell type identification. Hao et al. (2021)[28] provides high-resolution labels with 57 categories for PBMC datasets, encompassing all major and minor immune cell types. Our task is a multi-classification task that classifies observed cells into 57 categories, which poses a challenge since the subtypes may exhibit similar expression patterns. Figure

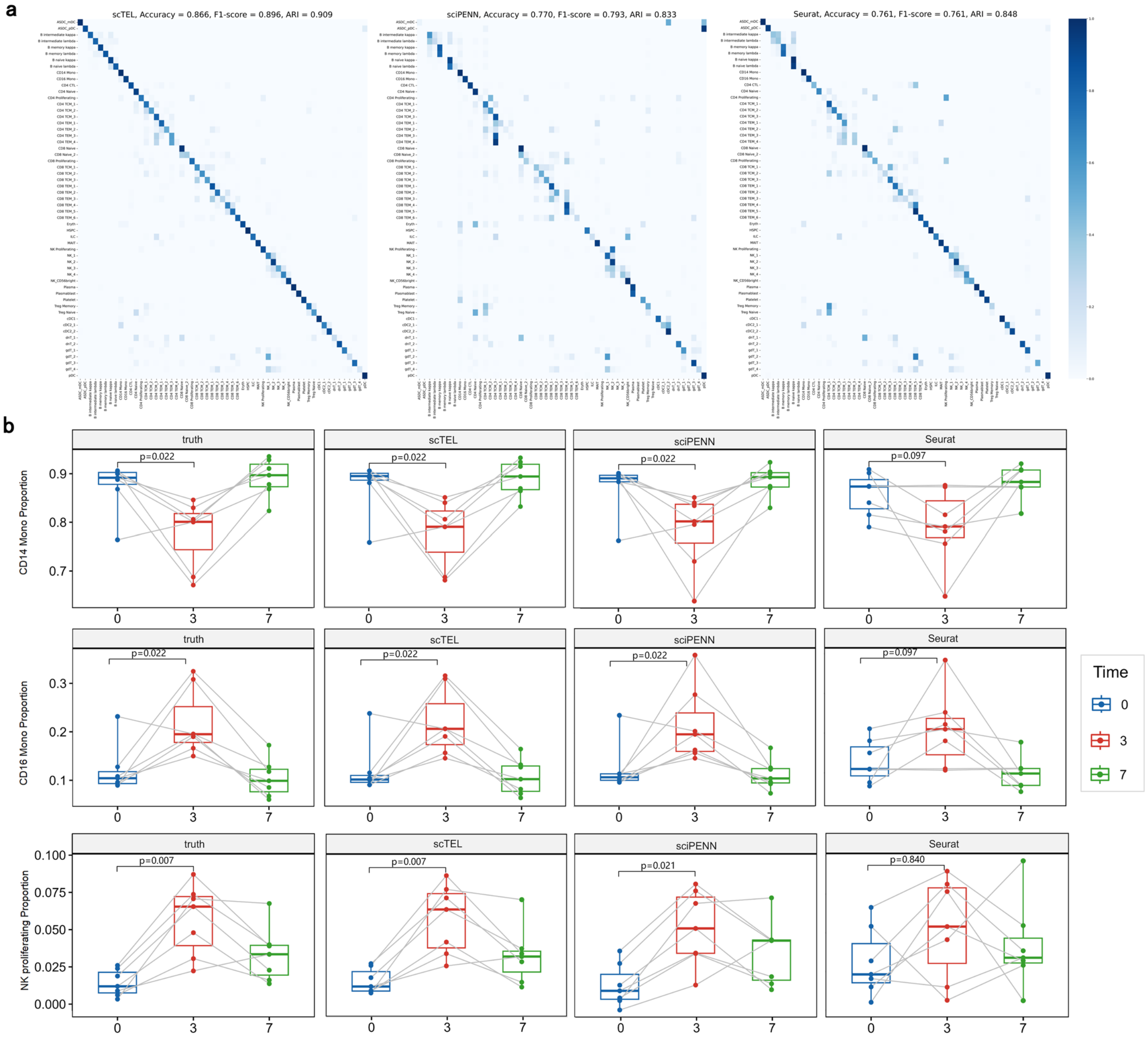

Fig. 6 | The performance of models for cell type identification. **a** The heatmap of the confusion matrix visualizes the classification performance, where the rows represent the true cell types and the columns represent the predicted cell types. All the three assessment metrics (accuracy, F1-score, and ARI) demonstrate that our scTEL model outperforms existing approaches. **b** The variation in the proportion of CD14 monocytes, CD16 monocytes, and NK proliferating at 3 time points in seven donors. The *p* values are reported from a paired Wilcoxon signed-rank test to compare the proportions of cells before vaccination and 3 days after vaccination.

6a present the confusion matrix, where the rows represent the true cell types and the columns represent the predicted cell types. Overall, scTEL achieves a significantly higher accuracy of 86.6% in cell type classification, outperforming sciPENN (77.0%) and Seurat (76.1%). Additionally, we assessed the classification performance using the F1-score and Adjusted Rand Index (ARI)[46], as shown in Fig. 6a. The results consistently indicate that our scTEL model surpasses existing approaches across all evaluation metrics.

Moreover, donors in the PBMC dataset received a vesicular stomatitis virus (VSV)-vectored HIV vaccine. Cell expression profiles were collected from patients at 3 time points: immediately before the vaccine, 3 days after the vaccine, and then 7 days after the vaccine. Hao et al. (2021)[28] reported that CD14 monocytes, CD16 monocytes, and NK proliferating exhibit distinct responses to the vaccine. Figure 6b illustrates the variation in the proportion of these three subtypes at different time points. Our scTEL model provides the closest predictions to the true pattern compared with sciPENN and Seurat.

Discussion

Life is a complex system. In modern molecular biology, according to the central dogma, genetic information is transcribed into RNA and then translated into proteins, which play a central role in shaping the diverse functions of cells and organisms. However, due to limitations in techniques and the high costs associated with experiments, research on proteomics is often challenging and constrained. These limitations hinder our comprehensive understanding of life activity as a whole.

In this paper, we propose an integrated framework, scTEL, to model RNA and protein expression jointly at the single-cell level. Leveraging CITE-seq datasets, our scTEL model establishes a mapping from sequenced RNA expression to the unobserved protein expression within the same cells. This advancement enables the prediction of protein expression using scRNA-seq data at a lower cost. In a unified framework, we achieve data integration, protein expression prediction, and cell type identification, forming a complete workflow for single-cell analysis. Through empirical validation on public datasets, we demonstrate that scTEL significantly outperforms traditional methods such as Seurat or totalVI in various tasks. Moreover, it exhibits superior performance compared to the RNN-based model sciPENN. The architecture of Transformer encoding layers and LSTM cells in scTEL significantly contributes to feature extraction in multi-omics data analysis.

In practice, scTEL still has several limitations. First, its sophisticated deep learning framework requires high computational resources, especially when handling large-scale single-cell datasets. Striking a balance between model performance and computational costs remains a challenging issue. Second, the training of the scTEL model relies on the access to reliable and well-annotated single-cell datasets, such as PBMC, which may limit its broader applicability.

Nowadays, integrated analysis of multi-omics data is becoming increasingly essential in cell biological research. It is crucial to analyze various related data types together to gain a comprehensive understanding of cellular processes. In the future work, we will consider to incorporate additional omics data, such as metabolomics and epigenetics, into our analysis framework.

Data availability

All datasets used in our study are from previously published studies. The PBMC dataset can be acquired through the Gene Expression Omnibus database (Accession number: GSE164378). The H1N1 dataset is published in[29] and can be acquired at https://doi.org/10.35092/yhjc.c.4753772. The Monocyte dataset is available at https://upenn.app.box.com/s/64c9fsex50g1bhv67893cpdg9c5jqjzo. The MALT dataset can be found at https://www.10xgenomics.com/resources/datasets/10-k-cells-from-a-malt-tumor-gene-expression-and-cell-surface-protein-3-standard-3-0-0.

Code availability

The code for scTEL has been uploaded to GitHub: https://github.com/142857cyy/scTEL.

Received: 13 August 2024; Accepted: 24 December 2024;
Published online: 02 January 2025

References

1. Vandereyken, K., Sifrim, A., Thienpont, B. & Voet, T. Methods and applications for single-cell and spatial multi-omics. *Nat. Rev. Genet.* **24**, 494–515 (2023).
2. Kolodziejczyk, A. A., Kim, J. K., Svensson, V., Marioni, J. C. & Teichmann, S. A. The technology and biology of single-cell RNA sequencing. *Mol. Cell* **58**, 610–620 (2015).
3. Grün, D. & van Oudenaarden, A. Design and analysis of single-cell sequencing experiments. *Cell* **163**, 799–810 (2015).
4. Badia-i Mompel, P. et al. Gene regulatory network inference in the era of single-cell multi-omics. *Nat. Rev. Genet.* **24**, 739–754 (2023).
5. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat. Methods* **14**, 865–868 (2017).
6. Maier, T., Güell, M. & Serrano, L. Correlation of mRNA and protein in complex biological samples. *FEBS Lett.* **583**, 3966–3973 (2009).
7. Hoernes, T. P., Hüttenhofer, A. & Erlacher, M. D. mRNA modifications: Dynamic regulators of gene expression? *RNA Biol.* **13**, 760–765 (2016).
8. Baysoy, A., Bai, Z., Satija, R. & Fan, R. The technological landscape and applications of single-cell multi-omics. *Nat. Rev. Mol. Cell Biol.* **24**, 695–713 (2023).
9. Bennett, H. M., Stephenson, W., Rose, C. M. & Darmanis, S. Single-cell proteomics enabled by next-generation sequencing or mass spectrometry. *Nat. Methods* **20**, 363–374 (2023).
10. Berridge, M. J. Unlocking the secrets of cell signaling. *Annu. Rev. Physiol.* **67**, 1–21 (2005).
11. Davis, D. M. Intercellular transfer of cell-surface proteins is common and can affect many stages of an immune response. *Nat. Rev. Immunol.* **7**, 238–243 (2007).
12. Su, Y. et al. Multi-omics resolves a sharp disease-state shift between mild and moderate COVID-19. *Cell* **183**, 1479–1495 (2020).
13. Revelo, X. S. et al. Cardiac resident macrophages prevent fibrosis and stimulate angiogenesis. *Circ. Res.* **129**, 1086–1101 (2021).
14. Wu, S. Z. et al. A single-cell and spatially resolved atlas of human breast cancers. *Nat. Genet.* **53**, 1334–1347 (2021).
15. Lake, B. B. et al. Integrative single-cell analysis of transcriptional and epigenetic states in the human adult brain. *Nat. Biotechnol.* **36**, 70–80 (2018).
16. Gawad, C., Koh, W. & Quake, S. R. Single-cell genome sequencing: Current state of the science. *Nat. Rev. Genet.* **17**, 175–188 (2016).
17. Yuan, M., Chen, L. & Deng, M. Clustering CITE-seq data with a canonical correlation-based deep learning method. *Front. Genet.* **13**, 977968 (2022).
18. Hao, Y. et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. *Nat. Biotechnol.* **42**, 293–304 (2024).
19. Gayoso, A. et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. *Nat. Methods* **18**, 272–282 (2021).
20. Tran, H. T. N. et al. A benchmark of batch-effect correction methods for single-cell RNA sequencing data. *Genome Biol.* **21**, 1–32 (2020).
21. Zhou, Z., Ye, C., Wang, J. & Zhang, N. R. Surface protein imputation from single cell transcriptomes by deep neural networks. *Nat. Commun.* **11**, 651 (2020).
22. Xu, J., Huang, D.-S. & Zhang, X. scmFormer integrates large-scale single-cell proteomics and transcriptomics data by multi-task Transformer. *Adv. Sci.* **11**, 2307835 (2024).
23. Lakkis, J. et al. A multi-use deep learning method for CITE-seq and single-cell RNA-seq data integration with cell surface protein prediction and imputation. *Nat. Mach. Intell.* **4**, 940–952 (2022).
24. Ribeiro, A. H., Tiels, K., Aguirre, L. A. & Schön, T. Beyond exploding and vanishing gradients: analysing RNN training using attractors and smoothness. In *International Conference on Artificial Intelligence and Statistics*, 2370–2380 (PMLR, 2020).

25. Yu, Y., Si, X., Hu, C. & Zhang, J. A review of recurrent neural networks: LSTM cells and network architectures. *Neural Comput.* **31**, 1235–1270 (2019).
26. Vaswani, A. et al. Attention is all you need. *Advances in Neural Information Processing Systems* **30** (2017).
27. Wu, T. et al. A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA J. Autom. Sin.* **10**, 1122–1136 (2023).
28. Hao, Y. et al. Integrated analysis of multimodal single-cell data. *Cell* **184**, 3573–3587 (2021).
29. Kotliarov, Y. et al. Broad immune activation underlies shared set point signatures for vaccine responsiveness in healthy individuals and disease activity in patients with lupus. *Nat. Med.* **26**, 618–629 (2020).
30. He, Z. et al. Mosaic integration and knowledge transfer of single-cell multimodal data with midas. *Nat. Biotechnol.* **42**, 1594–1605 (2024).
31. Zhou, S., Li, Y., Wu, W. & Li, L. scMMT: A multi-use deep learning approach for cell annotation, protein prediction and embedding in single-cell rna-seq data. *Brief. Bioinforma.* **25**, bbad523 (2024).
32. Yu, H., Zheng, Y. & Yang, X. scdm: A deep generative method for cell surface protein prediction with diffusion model. *J. Mol. Biol.* **436**, 168610 (2024).
33. Zheng, G. X. et al. Massively parallel digital transcriptional profiling of single cells. *Nat. Commun.* **8**, 14049 (2017).
34. Hafemeister, C. & Satija, R. Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression. *Genome Biol.* **20**, 296 (2019).
35. Wolf, F. A., Angerer, P. & Theis, F. J. Scanpy: Large-scale single-cell gene expression data analysis. *Genome Biol.* **19**, 1–5 (2018).
36. Szałata, A. et al. Transformers in single-cell omics: a review and new perspectives. *Nat. Methods* **21**, 1430–1443 (2024).
37. McInnes, L., Healy, J. & Melville, J. Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426 (2018).
38. Koenker, R. *Quantile regression* (Cambridge University Press, Cambridge, England, 2005).
39. Mao, A., Mohri, M. & Zhong, Y. Cross-entropy loss functions: Theoretical analysis and applications. In *International Conference on Machine Learning*, 23803–23828 (PMLR, 2023).
40. Zhang, F., Wu, Y. & Tian, W. A novel approach to remove the batch effect of single-cell data. *Cell Discov.* **5**, 46 (2019).
41. Rousseeuw, P. J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
42. Narasimhan, P. B., Marcovecchio, P., Hamers, A. A. & Hedrick, C. C. Nonclassical monocytes in health and disease. *Annu. Rev. Immunol.* **37**, 439–456 (2019).
43. Ravenhill, B. J., Soday, L., Houghton, J., Antrobus, R. & Weekes, M. P. Comprehensive cell surface proteomics defines markers of classical, intermediate and non-classical monocytes. *Sci. Rep.* **10**, 4560 (2020).
44. Siletti, K. et al. Transcriptomic diversity of cell types across the adult human brain. *Science* **382**, eadd7046 (2023).
45. Yao, Z. et al. A high-resolution transcriptomic and spatial atlas of cell types in the whole mouse brain. *Nature* **624**, 317–332 (2023).
46. Vinh, N. X., Epps, J. & Bailey, J. Information theoretic measures for clusterings comparison: is a correction for chance necessary? In *Proceedings of the 26th Annual International Conference on Machine Learning*, 1073–1080 (2009).

Acknowledgements

This research was funded by the National Natural Science Foundation of China [12401391, 12271217], the Natural Science Research General Project of Jiangsu Province [23KJB520003] and a scientific grant by Jiangsu Commission of Health [ZD2021009].

Author contributions

Conceptualization: Y.C. and C.W. Methodology: Y.C. and X.F. Data analysis: Y.C. and C.W. Writing-original draft: Y.C. and C.S. Writing-review and editing: X.F., Z.S., and C. W. Funding acquisition: Z.S. and C.W.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41540-024-00484-9.

Correspondence and requests for materials should be addressed to Chaojie Wang.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.