Genome Biology

RESEARCH

Open Access

A sequence knowledge-guided deep learning method for single-cell multi-omics translation

Mengyuan Zhao[1,2†], Jiawei Li[4†], Yanlin Jiang[5], Jiahui Yan[5], Xinyue Tang[7], Cheng Liang[6], Jijun Tang[1,2*] and Fei Guo[3*]

†Mengyuan Zhao and Jiawei Li contributed equally to this work.

*Correspondence: jj.tang@siat.ac.cn; guofei@csu.edu.cn

[1] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen 518055, China
[2] University of Chinese Academy of Sciences, Beijing 100190, China
[3] School of Computer Science and Engineering, Central South University, Changsha 410083, China
[4] College of Intelligence and Computing, Tianjin University, Tianjin 300350, China
[5] College of Engineering, Southern University of Science and Technology, Shenzhen 518055, China
[6] School of Information Science and Engineering, Shandong Normal University, Jinan 250358, China
[7] Shenzhen University of Advanced Technology, Shenzhen 518107, China

Abstract

Background: Analysis of proteins is key to understanding biological processes, disease pathogenesis, and advancing therapeutic development. However, proteome profiling remains significantly limited when compared to the exponential growth of single-cell RNA sequencing data, owing to technical challenges and prohibitive costs associated with large-scale protein detection. Recent advancements in multi-omics technologies have established essential connections between transcriptome and proteome layers, facilitating innovative computational approaches for predicting protein abundance based on transcriptome data.

Results: Here, we present scProTrans, an interpretable deep learning framework that synergizes sequence knowledge and multi-omics integration to achieve cross-omics translation in single-cell resolution. Our framework deciphers gene-protein associations through three innovative components: Firstly, a hierarchical attention mechanism that aligns gene/protein sequences with cellular contexts using CITE-seq training data; secondly a bidirectional encoder architecture implementing sequence-to-embedding-to-profile learning for modality translation; finally cell-specific associations capturing dynamic gene-protein interplay across heterogeneous cell populations. Extensive evaluations across 17 multi-omics datasets demonstrate that scProTrans surpasses state-of-the-art methods in single-cell protein abundance translation and enhances downstream analyses, including cell clustering, subtype identification, and biomarker discovery. scProTrans improves protein prediction accuracy and preserves low-abundance protein signals, two significant aspects of single-cell protein abundance translation. Additionally, scProTrans is extended to tri-omics scenarios (ATAC-RNA-protein) via modular encoder refactoring, achieving cross-modal prediction concordance comparable to experimental replication.

Conclusions: This work advances multi-omics integration by establishing a sequence-aware paradigm for cross-modal translation, overcoming key limitations in proteome data acquisition. This modular architecture and its zero-shot capability make it a versatile platform for emerging multi-modal single-cell technologies.

Keywords: Multi-omics, Single-cell sequencing, Zero-shot translation, Proteome profiling

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

Background

Single-cell transcriptome sequencing technology (scRNA-seq) has emerged as a highly potent tool for deciphering cell heterogeneity and regulatory networks of gene expression. It empowers us to characterize the landscape of gene expression at single-cell resolution, revealing specific transcriptional signatures of cell types, as well as the different states of cells during development, disease, and other processes [1–3]. However, it is crucial to note that proteins serve as the ultimate executors of gene expression. As fundamental biomacromolecules, proteins undertake a diverse range of core functions within cellular life activities, encompassing regulating gene expression, mediating signal transduction pathways, and catalyzing metabolic reactions. As direct implementers of cellular biological functions, proteins are closely related to enzyme activity regulation, post-translational modifications, and protein degradation processes [4, 5]. Consequently, an in-depth analysis of cellular mechanisms at the protein level provides a distinctive and pivotal perspective. Proteome offers distinct advantages over transcriptome by directly interrogating cellular activity functional molecules, circumventing confounding effects from transcriptional regulatory mechanisms. This methodological precision establishes its pivotal role in contemporary biomedical research, particularly in resolving tumor heterogeneity, advancing cancer immunotherapies, and accelerating therapeutic discovery. It provides unparalleled mechanistic insights into pathological processes while facilitating the identification of molecularly targeted treatment strategies [6, 7]. For example, Furnari et al. found that EGFRvIII protein exhibits significant expression differences among glioblastoma cells, which is closely related to drug resistance to EGFR-targeted therapy [8]. In summary, proteome research plays an indispensable role in augmenting and refining our comprehension of cellular life activities and disease mechanisms.

With the rapid development of multi-omics sequencing technology, CITE-seq enables simultaneous quantification of mRNA expression and surface protein epitopes in the same cell, thus revolutionizing cellular interrogation by bridging transcriptome and proteome dimensions at single-cell resolution. Joint analysis of multi-omics data empowers researchers to deconstruct cellular heterogeneity through complementary molecular lenses, constructing multidimensional cellular atlases that capture both transcriptional states and functional protein signatures [9–11]. However, since multi-omics sequencing experiments remain costly and complex endeavors, many studies conducted to date have utilized a relatively small number of cells. Furthermore, certain proteins present challenges for capture due to the lack of suitable antibodies. Compared to abundant scRNA-seq repositories, the scarcity of proteome data is particularly evident in clinical specimens and rare tissue types. To overcome these barriers, it is essential to develop predictive modeling frameworks that extract potential relationships between genes and proteins from existing multi-omics datasets. Such computational methodologies could establish paradigms for knowledge transfer, enabling the utilization of extensive scRNA-seq datasets for single-cell protein abundance inference across a variety of biological contexts.

Recent advances in multi-omics integration have spurred the development of computational frameworks capable of jointly analyzing heterogeneous molecular modalities. Notable algorithms in this domain include moETM [12], scVAEIT [13] and totalVI [14],

which employ variational autoencoders (VAEs) to model joint distributions of transcriptome and proteome data. Complementing these VAE-based methods, sciPENN and Seurat employs recurrent neural networks and weighted-nearest neighbor graphs to integrate CITE-seq datasets. While these methods demonstrate proficiency in aligning co-measured modalities or removing batch effects, few models have been explicitly designed for single-cell multi-omics translation, which is a critical capability for predicting unmeasured molecular profiles from available measurements.

Emerging solutions to the multi-omics translation challenge include scButterfly [15], which implements dual-aligned VAEs with modality-specific discriminators to achieve translation between transcriptome and proteome. The scTranslator framework [16] pioneers transformer-based architectures integrated with pretrained biological language models, enabling prediction of absent proteome profiles without additional training. However, current transcriptome-proteome translation approaches predominantly rely on paired multi-omics datasets, often overlooking valuable prior biological knowledge. This reliance results in dataset-specific performance limitations, restricting their generalizability. Moreover, these methods are confined to predicting proteins that are already profiled in training set, rendering them ineffective for proteins not profiled in existing multi-omics references and matched antibodies. Such constraints highlight the urgent need for more robust and versatile transcriptome-proteome translation models that can incorporate biological insights and extend beyond the boundaries of existing datasets.

To address these limitations, we've developed scProTrans, a modular deep learning framework based on cross-omics and cross-attention mechanisms. Using two omics-specialized encoders, this framework incorporates prior sequence knowledge of genes and proteins to predict protein profiles even without learning samples, reducing the impact of different datasets. With the cross-omics translation module, heterogeneous omics data are integrated into models of cell-specific gene-protein associations, translating transcriptome features into proteome features. As does the framework's core innovation, its dual-branch integration system captures fundamental gene-protein associations based on evolutionary-informed sequence embeddings, and has a context-aware transformer module that decodes single-cell molecular signatures by aligning omics-relevant representations across attention. The translation engine employs cross-attention mechanisms to resolve biological context-specific molecular mappings, revealing key regulatory drivers while maintaining single-cell resolution fidelity. Furthermore, the platform's modular design allows seamless expansion to tri-omics translation via plug-and-play encoder substitution and task-specific module augmentation, creating a scalable multi-omics synthesis system.

Results

Overview of scProTrans

In scProTrans, latent relations across omics modalities are learned by training single-omics query cells on paired multi-omics datasets and then transferred to single-omics query cells. Transcriptome and proteome reflect the pattern of gene and protein expression, respectively. As a result, scProTrans provides predictors with assistance from prior

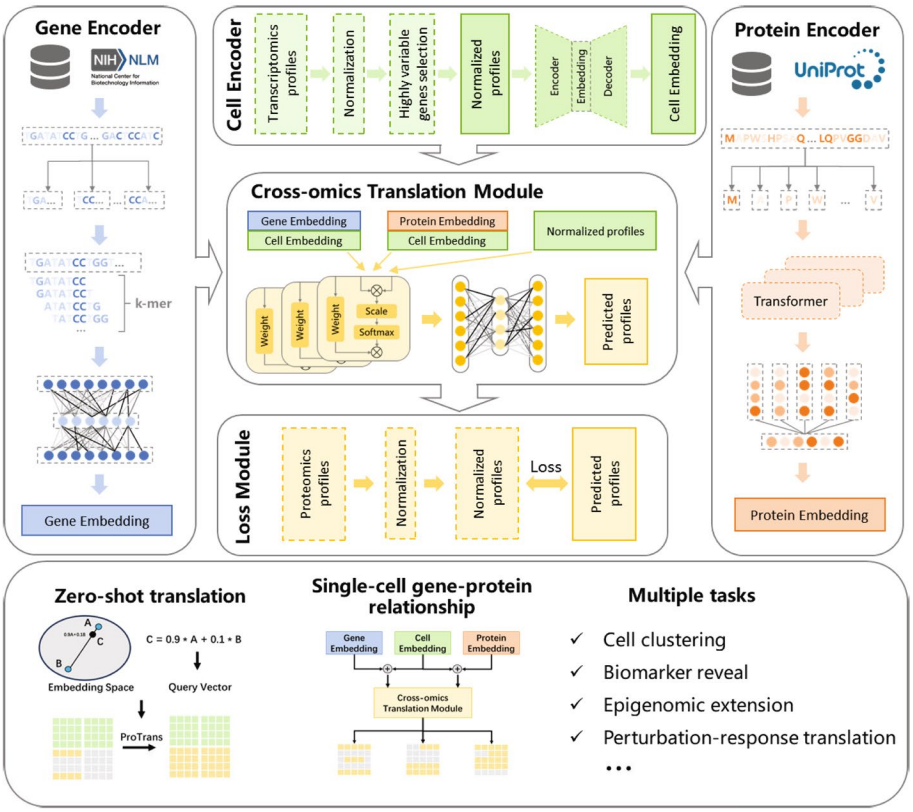


Fig. 1 The model architecture of scProTrans. A gene encoder creates a gene embedding for each gene from all human gene sequences using a combination of segmentation, k-mer computation, and artificial neural networks. A protein encoder produces a protein embedding for each protein from all human protein sequences using a protein language model based on transformers. A cell encoder aims to preprocess transcriptome data in order to obtain normalized profiles, followed by an autoencoder based deep learning algorithm to generate unique embeddings for each cell. In the cross-omics translation module, genes, proteins, cells, and normalized profiles are utilized as inputs to translate transcriptome profiles into proteome profiles using multi-head cross-attention techniques. In the loss module, the model is optimized iteratively by minimizing the difference between the normalized proteome data and the predicted results. scProTrans facilitates a range of downstream tasks, including zero-shot translation, single-cell gene-protein associations, cell clustering, biomarker discovery, perturbation-response translation, and tri-omics translation

biological knowledge, such as gene and protein sequences, as well as features extracted from transcriptome data.

The scProTrans architecture consists of five main modules (Fig. 1). Different biological perspectives are used by the gene encoder, the protein encoder, and the cell encoder to extract modality-specific features. The gene encoder integrates human gene sequences from NCBI database [17], processed by dna2vec [18] - a pre-trained k-mer embedded model - with sequence fragmentation and reassembly to produce fixed-length gene embeddings. Similarly, the protein encoder utilizes human protein sequences from UniProt database [19], encoded into embeddings via ProtT5 [20] to generate biologically meaningful representations. The cell encoder processes raw scRNA-seq data by normalizing expression profiles and identifying highly variable genes, and then applies a VAE framework to facilitate batch effects and derive cell-specific embeddings. Central to scProTrans is the cross-omics translation module, which implements a multi-head

cross-attention mechanism. Here, combined gene-cell embeddings serve as keys (K), combined protein-cell embeddings as queries (Q), and normalized expression profiles as values (V). Unique K-Q pairs are created for each cell through cell-specific embeddings, which allows us to examine gene-protein associations. Multi-layer artificial neural networks are used to convert the outputs of attention mechanisms into predictions of protein expression. Lastly, the loss module orchestrates the optimization process by integrating loss functions and early stopping criteria.

By incorporating sequence information, scProTrans not only transfers protein profiles from reference to query cells but also predicts the abundance of unavailable proteins (zero-shot translation), thus overcoming the limitation that CITE-seq datasets only contain cell surface proteins. scProTrans is capable of revealing single-cell gene-protein associations by combining omics' components with cell characteristics in embedding space. A variety of downstream tasks can be performed using scProTrans translation-based single-cell protein abundance data, including cellular clustering, perturbation-response analysis, and biomarker identification. Furthermore, scProTrans is able to handle tri-omics translation tasks as a result of its flexible and modular architecture.

Systematic evaluation of single-cell protein abundance translation

To evaluate scProTrans under real-world variability in multi-omics data analysis, we designed benchmark experiments spanning 14 publicly available datasets. These datasets encompass different batches, cell types, tissues, multi-omics sequencing technologies, and scales (cell numbers: 4,115–161,764). We compared scProTrans against six state-of-the-art methods: moETM [12], scVAEIT [13], totalVI [14], sciPENN [21], Seurat [22] and scButterfly [15]. In the following, we simulated four typical scenarios in a real-world situation to comprehensively assess the effectiveness of scProTrans.

Proteome translation on various benchmarking datasets

In this analysis, we harnessed 14 paired single-cell transcriptome-proteome datasets, featuring a diverse range of cell populations including bone marrow mononuclear cells (BMMCs), cord blood mononuclear cells (CBMCs), and peripheral blood mononuclear cells (PBMCs), among others. These datasets collectively capture the intricate cellular heterogeneity across multiple tissues and developmental stages, enabling a comprehensive exploration of the molecular landscape at the single-cell level. For in-depth details on the dataset characteristics, sequencing technologies, number of proteins, cell type and batch information, and associated metadata, we direct readers to Methods and Additional file 1: Table S1. For each dataset, 60% of cells were randomly selected for training, with the remaining 40% reserved for testing. The results, as shown in Fig. 2A, demonstrate that scProTrans outperforms existing methods in transcriptome-proteome translation. Specifically, scProTrans achieved lower mean squared error (MSE) and mean absolute error (MAE) values compared to competitors, with an average reduction of 0.04 in MSE and 0.02 in MAE. In terms of correlation, scProTrans exhibited a mean correlation coefficient of 92.2%, surpassing scButterfly and Seurat, which both achieved 91.9%. Moreover, scProTrans demonstrated superior stability, characterized by a minimal interquartile range (IQR = 0.04), indicating consistent performance in different datasets. To assess the robustness and reproducibility of our evaluation, we also conducted repeated

hold-out experiments, with detailed results provided in Additional file 2: Note S1, Fig. S1, and Additional file 1: Tables S2-S4.

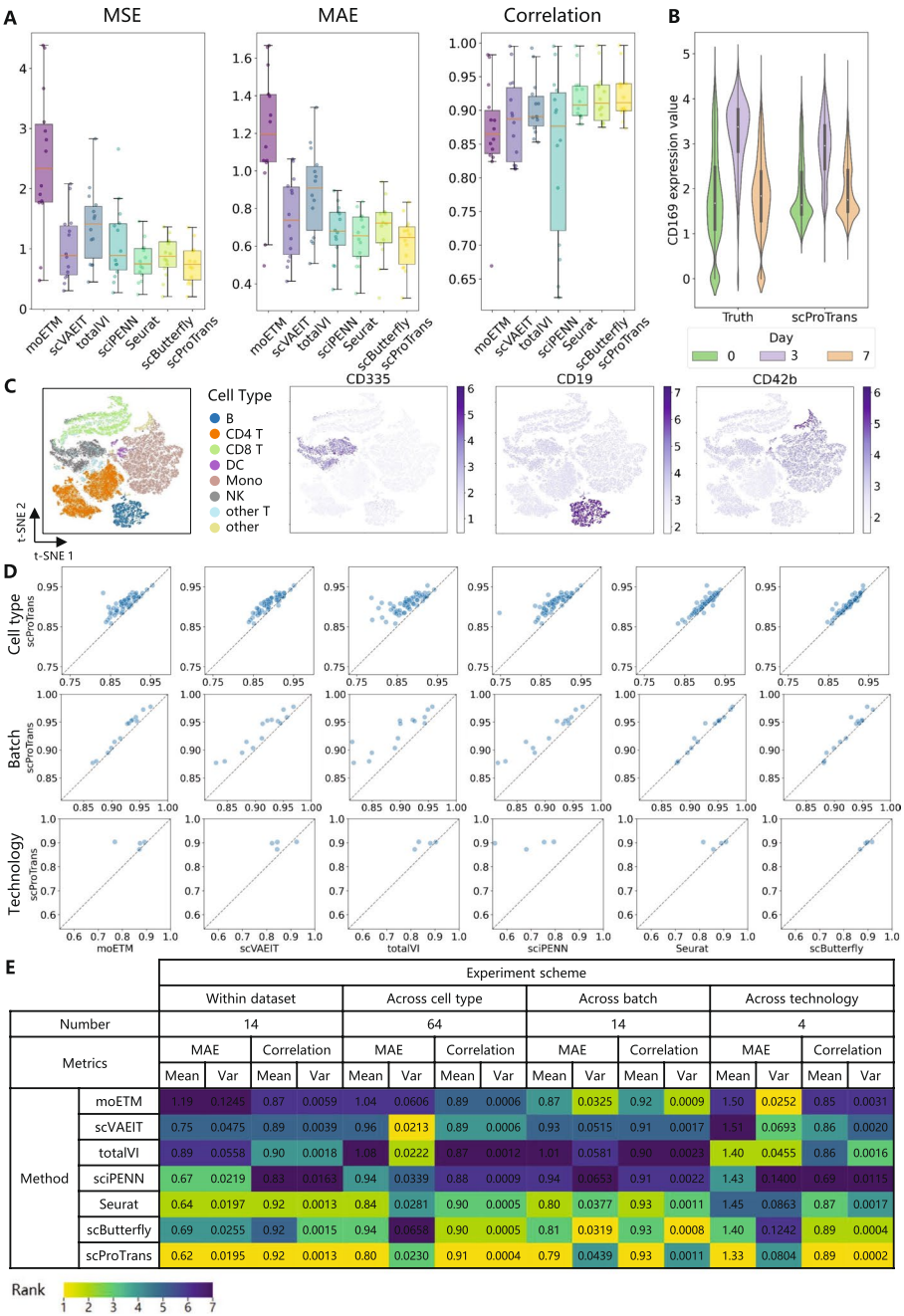
We further investigated the capacity of scProTrans to maintain biological characteristics regarding cell type and response to perturbation. Using the human PBMCs dataset (GSE164378) as a representative case, we leveraged t-Distributed Stochastic Neighbor Embedding (t-SNE) to visualize cell types based on CITE-seq measured protein expression and assess the scProTrans-translated profiles for individual proteins (Fig. 2C). Notably, scProTrans precisely recapitulated cell-type-specific protein expression signatures. For instance, CD335, a well-established NK cell marker [23], was accurately predicted to be highly expressed within the NK cell cluster and minimally detected in other cell populations. Similarly, the B cell marker CD19 and Mono marker CD42b showed distinct expression patterns within their respective cell clusters, underscoring scProTrans's ability to recover biologically relevant protein expression profiles. Moreover, scProTrans exhibited remarkable predictive power in tracking vaccine-induced protein expression dynamics (Fig. 2B, Additional file 2: Note S2 and Fig. S2). Analyzing the time-series proteome data from participants immunized with the VSV-vectored HIV vaccine in GSE164378, we focused on CD169 protein, a biomarker previously validated for its responsiveness to vaccination in CD14 Mono, CD16 Mono, and cDC2 cells [23]. scProTrans accurately predicted the temporal expression trajectory of CD169, closely mirroring the experimental observations with peak activation at day 3 post-vaccination followed by downregulation at day 7. These results not only highlight scProTrans's superiority in preserving biological characteristics but also its unique ability to capture stimulus-responsive protein expression changes. By enabling precise characterization of protein responses to external stimuli at single-cell resolution, scProTrans serves as a potent catalyst that facilitates the acceleration of vaccine development pipelines and the discovery of therapeutic targets.

Proteome translation across cell types

For certain cell types, it is difficult to obtain paired transcriptome-proteome data, but scRNA-seq data are widely available. To address this gap, we explored the feasibility of training models on multi-omics datasets and applying them across different cell types. Using three comprehensive datasets that collectively annotated 64 distinct cell

(See figure on next page.)

Fig. 2 Systematic comparison of proteome translation performance for scProTrans and state-of-the-art methods. **A** Box plots of MSE, MAE, and correlation between translated protein abundance and ground truth for each method. The lower and upper hinges indicate the first and third quartiles, respectively, and the center line denotes the median. Each point represents one dataset. **B** Violin plots showing CD169 protein expression before, three days, and seven days after vaccination with a VSV-vectored HIV vaccine. **C** t-SNE visualization of cells from the test set of the GSE164378 dataset, colored by cell types annotated by Seurat et al. [23]. Expression maps of marker proteins CD335, CD19, and CD42b show each cell colored according to protein abundance predicted by scProTrans. **D** Scatterplots comparing correlation values obtained by scProTrans and other methods across different experimental scenarios. Rows correspond to scenarios across cell types, batches, and technologies, respectively, and columns correspond to six state-of-the-art methods. Each point represents one experiment (e.g., a cell type, batch, or technology), with coordinates indicating the correlation between translated and ground-truth protein expression for the two methods. **E** Benchmark evaluation across scenarios involving targets within datasets, across cell types, batches, and technologies. Numbers indicate the number of experiments per scenario. Mean and Var denote the mean and variance of MAE or correlation across experiments. Colors represent method rankings for the same metric



E

		Experiment scheme															
		Within dataset				Across cell type				Across batch				Across technology			
Number		14				64				14				4			
Metrics		MAE		Correlation		MAE		Correlation		MAE		Correlation		MAE		Correlation	
		Mean	Var	Mean	Var	Mean	Var	Mean	Var	Mean	Var	Mean	Var	Mean	Var	Mean	Var
Method	moETM	1.19	0.1245	0.87	0.0059	1.04	0.0606	0.89	0.0006	0.87	0.0325	0.92	0.0009	1.50	0.0252	0.85	0.0031
	scVAEIT	0.75	0.0475	0.89	0.0039	0.96	0.0213	0.89	0.0006	0.93	0.0515	0.91	0.0017	1.51	0.0693	0.86	0.0020
	totalVI	0.89	0.0558	0.90	0.0018	1.08	0.0222	0.87	0.0012	1.01	0.0581	0.90	0.0023	1.40	0.0455	0.86	0.0016
	sciPENN	0.67	0.0219	0.83	0.0163	0.94	0.0339	0.88	0.0009	0.94	0.0653	0.91	0.0022	1.43	0.1400	0.69	0.0115
	Seurat	0.64	0.0197	0.92	0.0013	0.84	0.0281	0.90	0.0005	0.80	0.0377	0.93	0.0011	1.45	0.0863	0.87	0.0017
	scButterfly	0.69	0.0255	0.92	0.0015	0.94	0.0658	0.90	0.0005	0.81	0.0319	0.93	0.0008	1.40	0.1242	0.89	0.0004
	scProTrans	0.62	0.0195	0.92	0.0013	0.80	0.0230	0.91	0.0004	0.79	0.0439	0.93	0.0011	1.33	0.0804	0.89	0.0002

Rank 1 2 3 4 5 6 7

Fig. 2 (See legend on previous page.)

types, we conducted a leave-one-out cross-validation strategy. In this setup, each cell type was sequentially designated as the test set, while the remaining cell types constituted the training data. The 64 cell types corresponding to the 64 points shown in each scatterplot of cell type panel in Fig. 2D. scProTrans consistently outperformed state-of-the-art methods, with nearly all data points above the diagonal in comparative analyses (Fig. 2D). In these scatterplots, each point corresponds to one held-out cell type, with coordinates given by the correlation achieved by scProTrans and the corresponding baseline on the same test set. Additional file 2: Fig. S3A provides correlation metrics

obtained by each method on each cell type, clearly showing that scProTrans ranked first in almost all cases. Combined all cross cell type experiments, scProTrans possessed higher correlation and stronger stability than other methods (Additional file 2: Fig. S3B). This outcome underscores scProTrans's robustness and generalizability, establishing it as a reliable approach for cross cell type proteome prediction.

Proteome translation across batches

We evaluated the performance of scProTrans in cross batch scenarios, utilizing 14 batches from two datasets (Additional file 2: Fig. S4). For each dataset, we adopted a leave-one-batch-out evaluation strategy, where each batch was sequentially treated as the test set and all remaining batches were used for training. The 14 batches corresponding to the 14 points shown in each scatterplot of batch panel in Fig. 2D. In correlation scatterplots comparing scProTrans with moETM, scVAEIT, totalVI, and sciPENN, all points were located above the diagonal, indicating that scProTrans consistently achieved more realistic translations across all batches. When compared to Seurat and scButterfly, scProTrans demonstrated comparable performance in certain batches, while maintaining a clear advantage in others (Fig. 2D). Notably, although Seurat, scProTrans, and moETM showed competitive results in the GSE194122 dataset, Seurat exhibited limited generalizability in the GSE164372 dataset (Additional file 2: Fig. S4). In contrast, scProTrans emerged as the clear leader, achieving the highest median correlation (94.8%) across all batches. This superior performance highlights the reliability of scProTrans in handling batch-to-batch variation during multi-omics translations.

Proteome translation across sequencing technologies

The cross sequencing technology scenario evaluates scProTrans's effectiveness on two datasets encompassing three prevalent single-cell multi-omics technologies (CITE-seq, DOGMA-seq, and TEA-seq). As a result, we designed four pairwise experiments: CITE-DOGMA, DOGMA-CITE, CITE-TEA and TEA-CITE. As an example, CITE-DOGMA refers to training using CITE-seq data and testing using DOGMA-seq data. These experiments correspond to the four points shown in each scatterplot of technology panel in Fig. 2D. As illustrated in Additional file 2: Fig. S5, scProTrans, scVAEIT, and scButterfly demonstrated higher correlations under CITE-seq and DOGMA-seq conditions, whereas scProTrans, moETM, and totalVI excelled under CITE-seq and TEA-seq conditions. Overall, scProTrans exhibited the highest median correlation and lowest IQR compared to other methods. More detailed evaluation procedures for different translation scenarios are provided in Additional file 2: Note S3 and Fig. S6.

To comprehensively assess the robustness of different methods, we integrated the results from the four experimental strategies and conducted a systematic evaluation using mean absolute error (MAE) and correlation metrics, as illustrated in Fig. 2E. Among the evaluated approaches, Seurat, scButterfly, and scProTrans outperformed their counterparts, occupying the top three positions in most experiments. However, scButterfly's performance in cross-cell type scenarios was inconsistent, as evidenced by its MAE metric's variance ranking in last place. scProTrans emerged as the top-performing method, outshining Seurat, especially in cross-technology settings. Across 96 simulated experiments, scProTrans achieved an outstandingly low mean MAE of 0.88 and

high mean correlation of 91.4% between translated protein profiles and ground truth, which attested to scProTrans's proficiency in precisely translating protein abundance with minimal deviation and high reproducibility. The robustness of scProTrans can be attributed to its novel incorporation of gene and protein sequence knowledge. This multi-dimensional information effectively compensated for the biases stemming from distinct data origins, guaranteeing consistent translation performance of scProTrans across diverse experimental contexts.

Performance on cell clustering and subtype revelation

Cell clustering is a crucial step in processing single-cell data for downstream analyses. scProTrans enhanced cell type identification and clustering on all datasets annotated with cell type information. For each dataset, cells were randomly split into training (60%) and test (40%) sets. In this study, we examined clustering performance based on six metrics (Adjusted Mutual Information (AMI), Normalized Mutual Information (NMI), Rand Index (RI), Adjusted Rand Index (ARI), Homogeneity (HOM), and Completeness (V)), as applied to protein, RNA, and scProTrans-predicted profiles. scProTrans consistently outperformed raw RNA and protein data across all metrics, due to the cell encoder and cross-omics translation module, which overcome the high sparsity and dropout events (Fig. 3A, Additional file 2: Note S4, Fig. S7, and Additional file 1: Tables S5-S6). In GSE194122 and GSE164378, cell clustering based on protein profiles tends to achieve better results than RNA profiles, as a result of the high dimensionality and sparseness of large-scale transcriptome datasets. Visualization of clustering results further highlights scProTrans's advantages (Fig. 3B and Additional file 2: Fig. S8A). In GSE164378, RNA-based dimensionality reduction failed to distinguish CD4 Naïve, CD4 TCM, and CD8 Naïve T cells, while protein-based clustering partially resolved CD8 Naïve from CD4 subsets but conflated it with CD8 TCM and MAIT cells. In contrast, scProTrans-predicted protein profiles clearly delineated CD8 Naïve T cells and separated CD4 Naïve from CD4 TCM populations, demonstrating superior resolution for fine-grained cell type identification. An highlighted version of this visualization is provided in Additional file 2: Fig. S9.

On the basis of the translated protein profiles, we focused on 3 subpopulations that belong to the same batch and type but exhibit spatial separation (Fig. 3B and Additional file 2: Fig. S8C). For each subpopulation, we employed the Wilcoxon rank-sum test in Scanpy [24] to identify the top 50 differentially expressed proteins (DEPs) (Additional file 2: Fig. S8B). Notably, no overlapping DEPs were observed among the three subpopulations (Fig. 3C), indicating distinct expression patterns. As illustrated in Fig. 3D, the elevated expression of CD49a in the bottom cluster suggests its potential as a subtype-specific marker for CD14 monocytes, a hypothesis supported by Martrus et al. [25]. CD142 is highly expressed in the upper-left cells, indicating their strong association with viral infections [26, 27]. Similarly, CD169 expression in the upper-right cluster correlates with activated monocyte phenotypes described in prior studies [28]. Collectively, scProTrans-translated proteome profiles not only characterize cell types, but also preserve subtle differences within the same type, contributing to subtype identification and biomarker discovery.

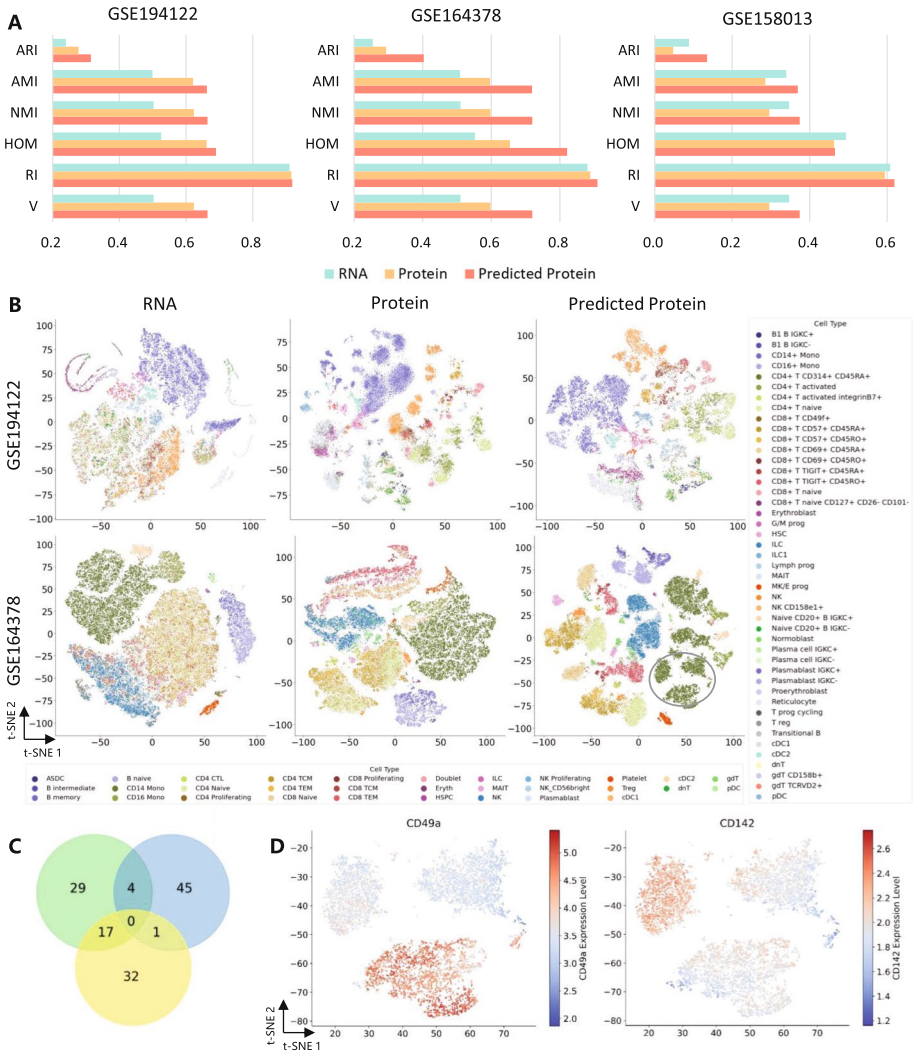


Fig. 3 Single-cell protein abundance translation assists in cell clustering and subpopulation identification. **A** Bar charts compare 6 clustering metrics calculated based on RNA expression, protein expression and predicted protein expression on 3 datasets. **B** Scatter plots visualization of t-SNE distribution based on RNA expression, protein expression, and scProTrans-predicted protein expression on different datasets. Scatter colors correspond to cell types. GSE194122 contains 45 cell types, and the legend is shown on the right. GSE164378 contains 31 cell types, and the legend is shown below. The 3 clusters within the ellipse are CD14 Mono subpopulations from scProTrans. The scatter plot corresponding to GSE158013 is provided in Additional file 2: Fig. S8A. **C** Venn diagram displaying the overlap among the top 50 DEPs from each subpopulation. **D** Predicted protein expression heatmap for CD49a and CD142 (also see CD196 in Additional file 2: Fig. S8D)

Discovering single-cell gene-protein associations via cross-omics translation module

Using a multi-head cross-attention mechanism, the cross-omics translation module quantifies single-cell gene-protein associations via attention matrices, bridging transcriptome and proteome modalities. To assess the biological significance of these relationships, we analyzed single-cell attention matrices within a cell type to produce cell-type-specific attention matrices. In GSE164378, we analyzed such matrices for B cells and CD4 T cells. Considering the genes and proteins highlighted by scProTrans, we collected marker genes corresponding to B cells and CD4 T cells from the CellMarker2.0

database [29]. As shown in Fig. 4A, B, genes with high attention weights correspond to established markers for various cell types. While most proteins exhibited uniform attention distributions, distinct patterns emerged for specific markers. For example, *BLNK* (Fig. 4A) showed low global attention but selectively assigned high values to CD21 and CD22. These proteins are related to B cell markers identified by Hao et al. [22]. They demonstrate that scProTrans preserves biological specificity by dynamically weighting marker-associated genes and proteins, thereby minimizing noise interference and allowing accurate cross-omics information transfer.

Marker gene identification typically relies on differential expression statistics, i.e., fold change. In contrast, the cross-omics translation module reveals marker genes from attention scores. Figure 4C, D, Additional file 2: Note S5 and Fig. S10 explore the association between fold change, attention scores and marker genes. In B cells, marker genes predominantly localize in the upper-right quadrant, while non-marker genes in CD4 T cells cluster in the lower-left quadrant. This spatial distribution highlights that attention scores contribute to locating marker genes. Quantitatively, combining attention scores with fold change improved marker gene determination precision by 34% for B cells and 21% for CD4 T cells. Furthermore, the inclusion of cell embeddings in the cross-omics translation module enables single-cell resolution analysis of gene-protein associations. Figure 4E intuitively visualizes them for two individual B cells, revealing shared patterns (e.g., consistently high-weight genes across proteins) alongside cell-specific variations. Their difference heatmap highlights weight disparities in specific genes and proteins, reflecting cellular heterogeneity and providing a foundation for single-cell level functional analyses.

Inferring unavailable protein expression profiles via zero-shot translation mechanism

Currently, cross-omics translation methods can only deduce the proteins observed in training sets. In spite of this, CITE-seq only detects cell surface proteins; therefore, more proteins cannot be measured directly or predicted due to a lack of matched multi-omics references. As opposed to other tools, scProTrans makes it possible via zero-shot mechanism. By leveraging sequence-derived biological priors, scProTrans is able to generalize even without experimental measurements to proteins of interest.

scProTrans's zero-shot translation performance for four proteins excluded from the training set, is illustrated in Fig. 5A. CD98 and CD274, which exhibit cluster-specific expression patterns, are accurately reconstructed by scProTrans. Similarly, the homogeneous distributions of CD144 and CD324 are faithfully reproduced, with MSE values of 0.51 and 0.87, respectively. More challenging, we also performed zero-shot translation for non-CD family proteins (CCR10, TIGIT, and XCR1). The predicted and true expression values are compared in Additional file 2: Fig. S11, with correlations of 89%, 77%, and 75%, respectively. To elucidate the zero-shot mechanism, we analyzed CD324 and its closest embedded neighbor from the training set (CD325). It means that scProTrans utilized CD325 as a reference for translating the unavailable protein CD324, which was treated as a query. Using CLUSTALW [30], sequence alignment revealed high sequence conservation and quality for the two proteins, with identical amino acids at most sites (Fig. 5C). The complete results are available in Additional file 2: Fig. S12. As a result of similar sequences, they attained an embedding similarity of 96.4%, with consistent

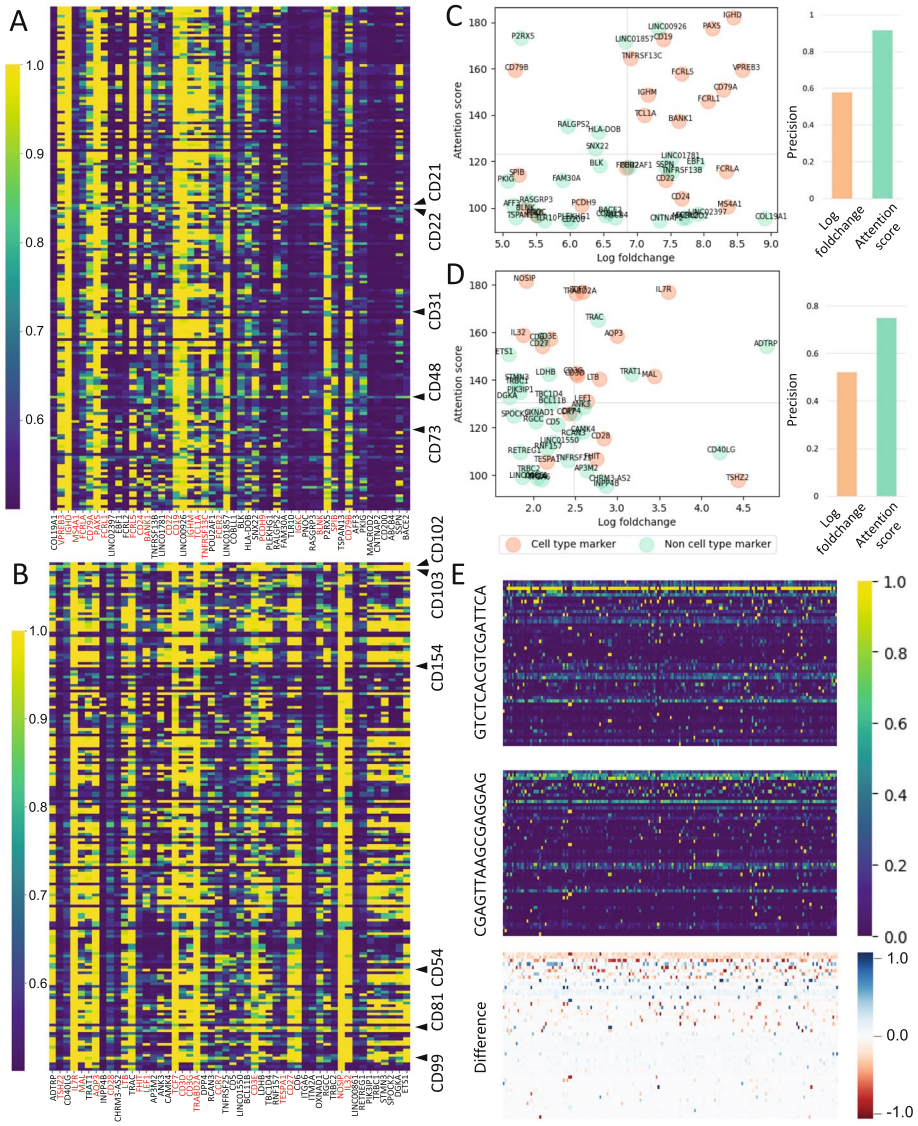


Fig. 4 Revealing biomarkers from the cross-omics translation module. **A** Heatmap of the gene-protein attention matrix of B cells. **B** Heatmap of the gene-protein attention matrix of CD4 T cells. Columns are the top 50 differentially expressed genes (DEGs), and in red font are marker genes. Rows are proteins, and those labeled on the right are cell type marker proteins. **C** The scatter plots of DEGs for B cells. **D** The scatter plot of DEGs for CD4 T cells. The x-axis is $Log_2(foldchange)$ and the y-axis is the attention score from scProTrans. The orange dots are cell type marker genes and correspond to red fonts in the heatmaps. The gray lines indicate the mean values for all genes. Bar charts display the precision of determining marker genes based on fold change or combined fold change and attention score. **E** Cell-specific gene-protein attention matrices. The top 2 heatmaps correspond to different cells, and the bottom heatmap visualizes their differences

high- and low-value sites (Fig. 5D). From the CITE-seq datasets, CD324 and CD325 also exhibit similar expression profiles, which explains why scProTrans enables zero-shot translation by integrating sequence information (Fig. 5B). This exemplifies the model's capacity to leverage sequence similarity for zero-shot translation, offering a novel approach to expand the coverage of multi-omics analysis beyond proteins with existing experimental data.

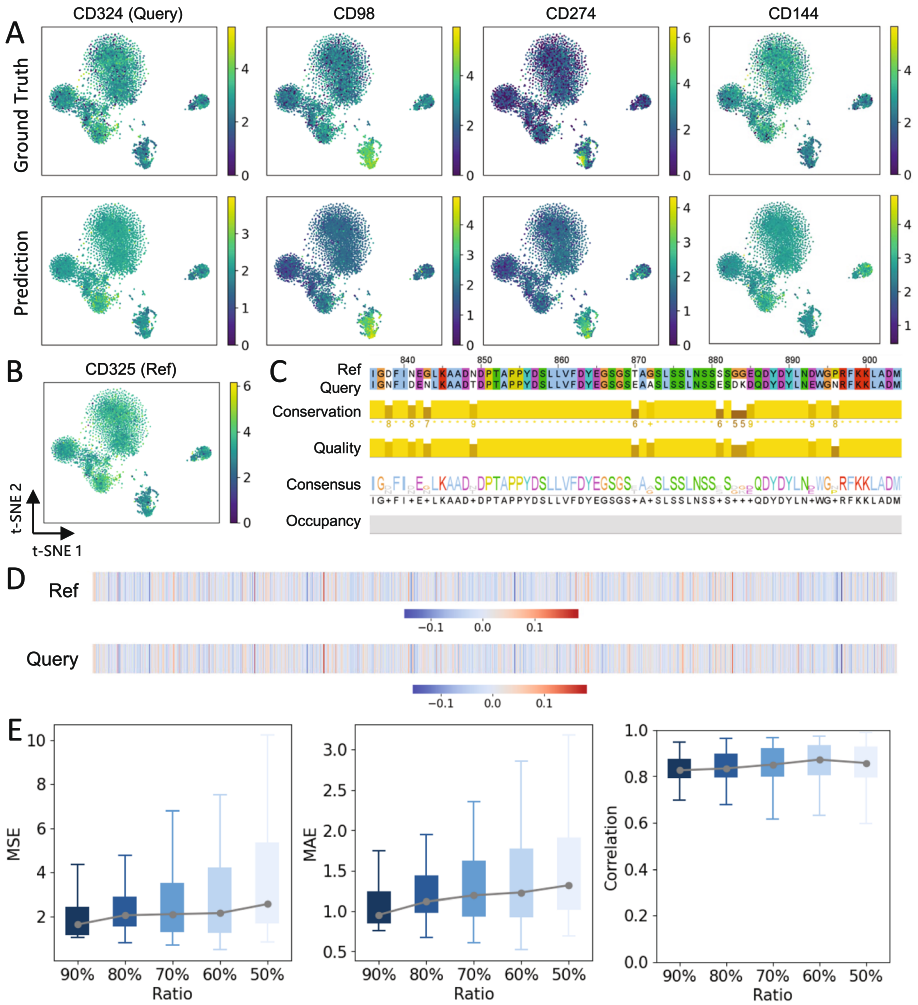


Fig. 5 Inferring novel protein expression profiles with zero-shot learning. **A** Scatterplots of ground truth and predicted values for 4 proteins excluded from the training set. **B** The scatterplot of ground truth profile for CD325 involved in the training set. **C** Sequence alignment analysis of reference and query proteins visualized by Jalview et al. [31]. Conservation serves as a quantitative numerical indicator that reflects the preservation of physical and chemical properties. Quality denotes the likelihood of observing mutations within a specific column. Consensus indicates the percentage of different residues present in each column, while occupancy represents the total number of residues in each column. Both consensus and occupancy provide insights into the diversity and abundance of residues across columns. **D** The protein embeddings of reference and query. Colors from blue to red indicate feature values from low to high. **E** Effect of training set ratio on zero-shot translation performance of scProTrans. The line connects mean values under different ratios

We further quantified the impact of training set size on zero-shot performance by systematically varying the proportion of proteins designated for inclusion in the training set (Fig. 5E). When reducing the proportion of proteins in the training set from 90% to 50%, the translation error for unseen proteins increases modestly, with MSE rising by 0.9 and MAE by 0.4, while the correlation between predictions and ground truth consistently remains above 80%. It demonstrates that scProTrans is robust to variations in the training ratio, and scaling to larger multi-omics datasets should further enhance its capabilities. As an additional biologically meaningful evaluation beyond error-based metrics, we further assessed zero-shot translation performance by comparing the overlap of highly

variable proteins (HVPs) identified from scProTrans predictions and from the ground truth across different train/test splits (Additional file 2: Note S6). Consistently high overlap ratios were observed, particularly when larger proportions of proteins were included in the training set, indicating that scProTrans preserves biologically relevant protein expression patterns under zero-shot settings.

Proteomic perturbation response analysis Via scProTrans

With the rapid accumulation of single-cell data, it is critical to explore the response of single cells to various types of perturbations, such as drug treatment and viral infection [32–35]. Such studies provide insights into molecular mechanisms and reveal the heterogeneity of disturbance responses among different cell types. Benefiting from the wide application of scRNA-seq, most available data and analysis regarding perturbation-response focus on transcriptome [36], which can be extended to proteome by scProTrans.

After training on a large-scale PBMCs multi-omics dataset (GSE164378), scProTrans predicted protein expression profiles for GSM2560248 and GSM2560249. They consist of scRNA-seq data from PBMCs of eight lupus patients, with control and interferon-beta-stimulated (IFN-β) groups across seven cell types [37]. Using translated single-cell protein abundance, differential expression analysis identified 33 DEPs (28 down-regulated, 5 up-regulated) between stimulated and control groups (first plot of Fig. 6A). Cell-type-specific volcano plots revealed distinct responses to IFN-β. CD14+ Mono and FCGR3A+ Mono exhibited the strongest responses, with over 90 DEPs and high statistical significance ($-\log_{10}(P-Value) > 60$). In contrast, other cell types showed fewer DEPs (<60) and lower significance ($-\log_{10}(P-Value) \approx 30$). These findings, consistent with prior studies [38–40], demonstrate that scProTrans-predicted proteome profiles capture both global perturbation effects and cell-type-specific responses, enabling deeper mechanistic insights into cellular behavior under perturbation.

To further characterize the functional implications of DEPs in CD14+ Mono and FCGR3A+ Mono, we performed Gene Ontology (GO) and pathway enrichment analyses (Fig. 6B). Both cell types exhibited GO enrichment related to cell adhesion, immune regulation, and receptor activity, consistent with priori conclusions [41–43]. In terms of biological processes, most DEPs aggregated to the immune response as expected. In contrast, FCGR3A+ Mono was enriched for cell-cell adhesion, whereas CD14+ Mono focused on leukocyte cell-cell adhesion, suggesting their different roles in the process of immune response. Pathway analysis highlighted key processes, including hematopoietic cells, cell migration, and immunodeficiency. Notably, the PI3K-Akt signaling pathway in Fig. 6B attests to the impact of dysregulation in PI3K-Akt pathway on lupus, with Iacoba et al. also suggesting it as a potential therapeutic target [44–46]. Consequently, the generation of perturbation-response matched proteome data by scProTrans facilitates the analysis of diverse cellular states and unveiling disease-associated pathways.

To directly evaluate whether scProTrans can recover perturbation-induced proteome data, we conducted an additional analysis using independent CITE-seq datasets, which

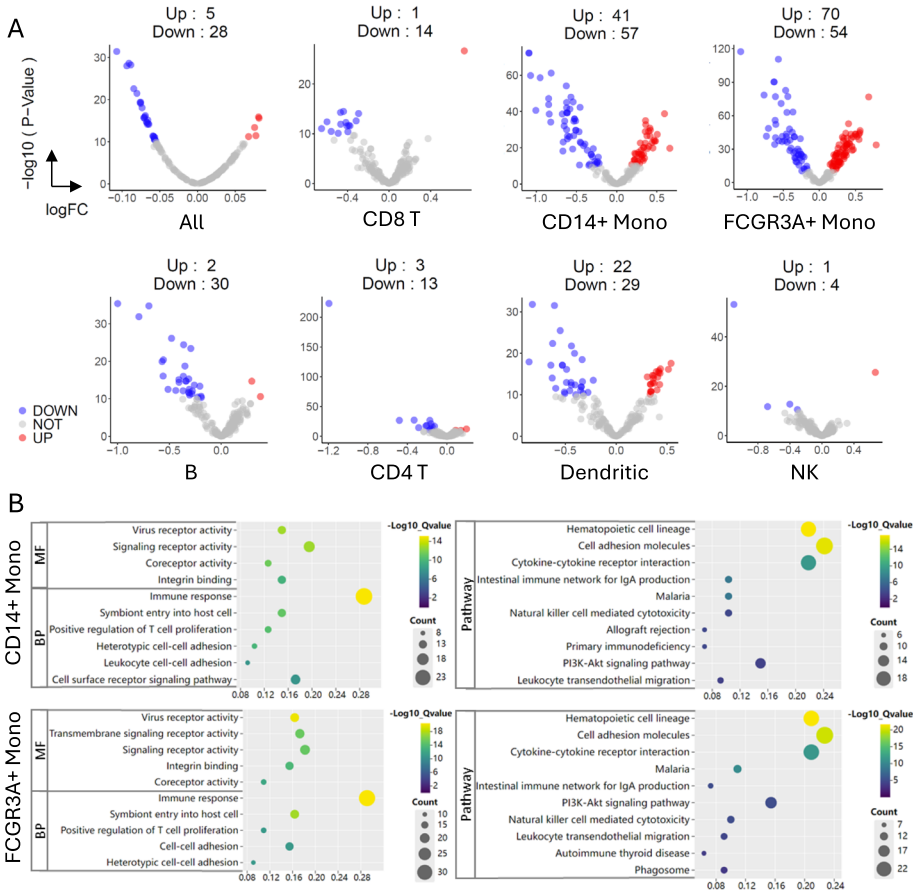


Fig. 6 Proteome-based perturbation response analysis. **A** Volcano diagrams of differentially expressed proteins for different cell types. The first plot shows differently expressed proteins for all cells, and the other plots are cell-type-specific differently expressed proteins. Blue and red points represent down- and up-regulated proteins, respectively. **B** GO enrichment and pathway enrichment for differentially expressed proteins of CD14+ Mono and FCGR3A+ Mono cells. BP and MF are abbreviations for biological processes and molecular functions. Count is the number of proteins enriched in each item

includes both control cells and cells stimulated IL2 and anti-CD3/CD28 (GSM4732113 and GSM4732115). They provide paired transcriptomic and proteomic measurements, enabling a direct assessment of protein-level response prediction. To enable translation from control transcriptome to perturbed proteome, cells from the control transcriptomic data and the perturbed proteomic data were separately split into training and testing sets with a 6:4 ratio. scProTrans was trained on the training set to learn the transcriptome-to-proteome translation mechanism while incorporating perturbation response patterns. The trained model was then applied to the transcriptome of control cells in the test set to predict perturbed protein expression profiles, which were quantitatively compared with the observed proteomic measurements of perturbed cells.

As shown in Additional file 2: Fig. S13A, the predicted and observed mean protein expression levels after stimulation exhibit strong agreement across all proteins, with most points closely distributed along the diagonal. Quantitatively, scProTrans achieves an MSE of 0.91, an MAE of 0.71, and a correlation of 88.86%, indicating accurate recovery of global perturbation-induced protein expression changes. We further examined

proteins that are differentially expressed in response to perturbation. Based on the observed proteomic profiles, the top-10 DEPs were identified using scanpy and visualized in Additional file 2: Fig. S13B, while the top-10 DEPs derived from the predicted proteomic profiles are shown in Additional file 2: Fig. S13C. Seven proteins overlap between the predicted and observed DEPs, demonstrating that scProTrans recovers a substantial fraction of perturbation-responsive proteins.

Beyond high overlap of DEPs, scProTrans also accurately captures the direction and relative magnitude of perturbation-induced expression changes. As illustrated in Additional file 2: Fig. S13D, all top-10 DEPs exhibit increased expression following perturbation, with varying effect sizes, and the predicted changes closely match the observed responses, including strong responders such as CD69 and weaker responders such as CD2. Finally, extending the analysis beyond the top-10 DEPs, we quantified the overlap between predicted and observed DEPs across increasing top-k thresholds. The overlap ratio consistently exceeds 70% and increases with larger k values (Additional file 2: Fig. S13E). In summary, these results demonstrate that scProTrans not only performs accurate cross-omics translation but also effectively captures perturbation-induced protein response patterns, supporting its applicability to perturbation-response biological analysis.

Extending scProTrans to additional omics modalities

To evaluate the cross-omics scalability of scProTrans, we extended the framework from its original RNA–protein translation task to multiple other omics modalities. Beyond transcriptome and proteome, emerging multi-omics technologies, such as DOGMA-seq and TEA-seq, enable simultaneous profiling of epigenome, transcriptome and proteome data within single cells [47, 48]. Leveraging matched tri-omics datasets, we extended scProTrans to translate across three omics. We evaluated scProTrans using five datasets with matched single-cell epigenome, transcriptome and proteome data.

For epigenome-to-proteome translation, we randomly partitioned ATAC and ADT data into training (60%) and test (40%) sets, replacing the gene encoder with a peak encoder to process the scATAC-seq data. Figure 7B illustrates the strong agreement between predicted and experimentally measured protein expression levels across all datasets, with most points aligned along the diagonal. Quantitative analysis (Fig. 7A) revealed high consistency, with mean correlations of 88.1%, 86.5%, 86.9%, 86.4%, and 86.7% across five datasets, respectively. The minimal variation (<1.7%) in performance underscores scProTrans's robustness and generalizability for epigenome-to-proteome translation analyses, which highlights its potential for integrated multi-omics analysis.

For epigenome-to-transcriptome translation, we adapted scProTrans by replacing gene and protein encoders with peak and gene encoders. Training and test sets were partitioned as described for epigenome-to-proteome translation. To evaluate performance, we calculated expression variance for each gene across test cells, selecting *LYN* (highest variance) and *RFX4* (lowest variance) as representative examples. Figure 7C compares the predicted and true profiles for these genes, demonstrating scProTrans's ability to accurately reconstruct distinct expression patterns. We further analyzed *RORA* and *FCRL1*, which are representative genes featuring differential expression in specific cell populations. In agreement with their ground truth profiles, translated *RORA* is highly

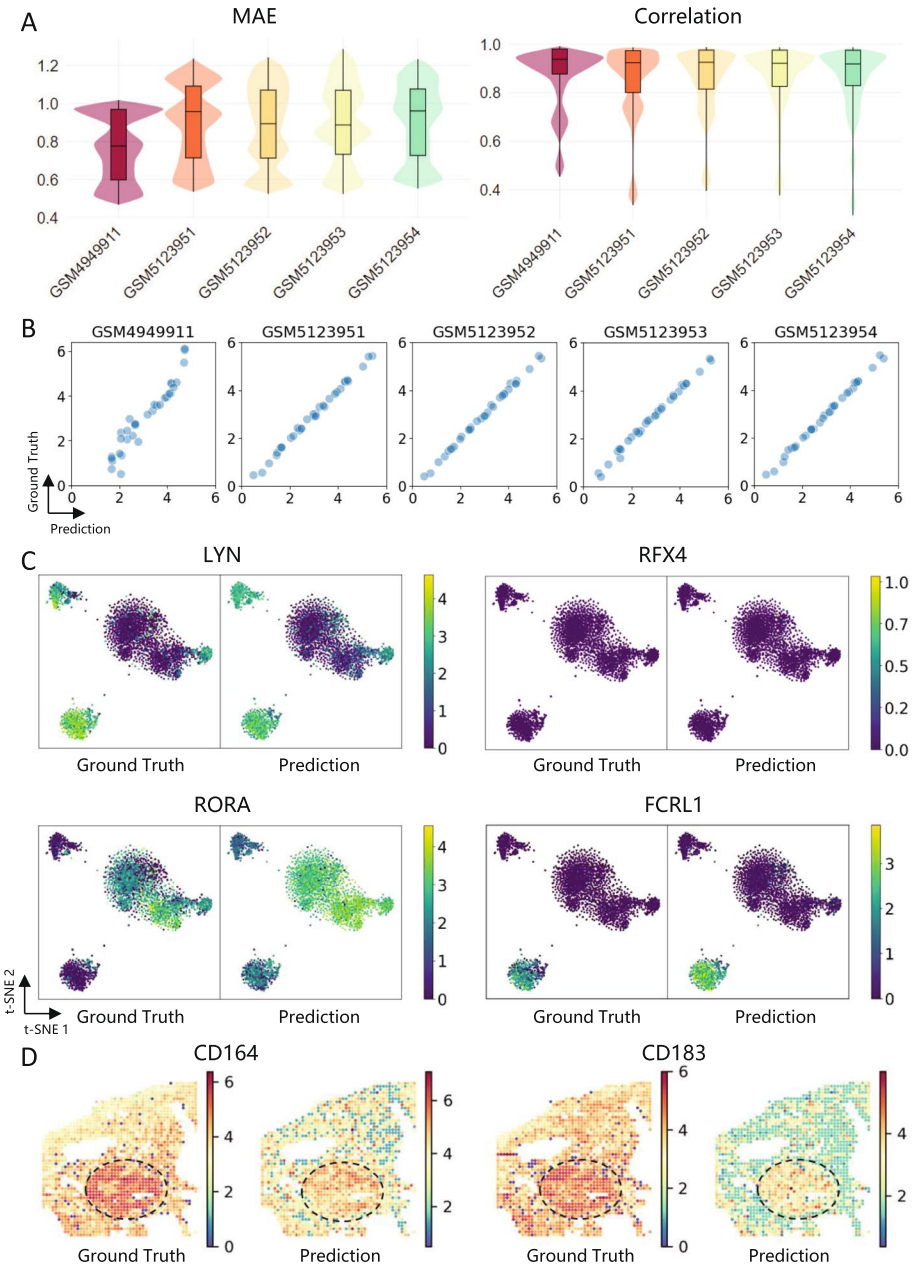


Fig. 7 Extending scProTrans to tri-omics translation. **A** MAE and the correlation between ground truth and translated protein expression values. The violin plot reflects not only the median, first and third quartiles but also the overall distribution of the corresponding metrics. **B** Performance of epigenome-to-proteome translation on multi-omics datasets including GSM4949911, GSM5123951, GSM5123952, GSM6123953 and GSM5123954. The x-axis of the scatterplot represents predicted values, the y-axis represents ground truth, and each point represents a protein. For each protein, the mean expression value across all cells serves as its expression value. **C** Scatterplots of epigenome-to-transcriptome translation results for genes including *LYN*, *RFX4*, *RORA* and *FCRL1*. For each scatterplot, the ground truth are shown on the left and scProTrans's predictions on the right. Each point represents a cell and color represents expression values. **D** Heatmaps of ground truth and predicted spatial expression of CD164 and CD183 protein. The dotted elliptical areas denote highly consistent areas

expressed only in the right cluster. *FCRL1* is highly induced in the lower left cluster. According to these results, scProTrans is capable of inferring differential expression patterns in transcriptome using epigenome information.

As spatial protein measurements remain scarce, costly, and technically challenging, translating spatial proteome (SP) from spatial transcriptome (ST) via scProTrans provides a feasible approach. To assess scProTrans in this context, we collected paired spatial transcriptomics and spatial proteomics data and extended the model to the spatial omics setting. Specifically, we used a human skin dataset from Liu et al. [49], comprising 1,691 spatial spots, 283 proteins, and 15,486 genes. We applied scProTrans to the spatial transcriptomics profiles and performed zero-shot prediction of the corresponding spatial proteomics measurements.

To evaluate the translation performance, we removed CD164 and CD183 from the training set. CD164 is closely associated with multiple skin disorders, including mycosis fungoides, Sézary syndrome, atopic dermatitis, and psoriasis, and has been proposed as a potential biomarker for malignant CD4 T cells [50]. scProTrans successfully recovered the spatial expression pattern of CD164 from spatial transcriptome. The predicted spatial proteome closely resembled the ground truth, capturing the characteristic ellipse-shaped region of high expression (Fig. 7D). The prediction achieved a MAE of 1.15 and a correlation of 93% compared to the measured proteomics data. CD183, a chemokine receptor implicated in transplant rejection and identified as a skin-homing marker in AML, was also evaluated under the zero-shot setting [51]. Although scProTrans slightly underestimated the overall expression magnitude of CD183, it accurately recovered its spatial distribution, again showing a consistent ellipse-shaped local high-expression region similar to the ground truth. The prediction yielded an MAE of 1.81 and a correlation of 90%. Overall, these case studies demonstrate that scProTrans can be effectively applied to spatial transcriptomics data to infer spatial proteomics signals, offering a practical and cost-efficient solution for spatial proteome profiling.

Discussion

scProTrans is a deep learning framework designed to translate proteome data from scRNA-seq datasets, enabling zero-shot translation while uncovering cell-specific gene-protein associations. By incorporating prior sequence knowledge, scProTrans integrates cell-specific and omics-specific features to explicitly learn the gene-protein associations at single-cell resolution. Systematic evaluations across 16 datasets—spanning diverse batches, tissues, and technologies—demonstrate scProTrans's superior performance over state-of-the-art methods. Beyond multi-omics translation, scProTrans has proven advantageous for diverse single-cell analyses, including cell clustering, subtype identification, biomarker discovery, and perturbation-response analysis.

In contrast to existing proteome prediction tools, scProTrans leverages prior sequence information, reducing dependence on specific datasets and enhancing its robustness. Furthermore, for proteins undetectable by multi-omics sequencing technologies, scProTrans utilizes sequence knowledge to predict the corresponding profiles. This zero-shot translation strategy is not confined to proteome; it can be applied to other omics, addressing the data scarcity caused by experimental constraints. Additionally, the

cross-omics translation module provides cell-specific gene-protein associations, which empower us to identify biomarkers while exploring associations of different omics at different resolutions including single cell, cell type, and tissue.

As a flexible and scalable tool, scProTrans has been successfully applied to tri-omics datasets, where sequence information corresponding to genes, proteins and peaks is utilized. In the future, one possible improvement is to integrate more biological knowledge, such as transcription factors and regulatory relationships. Another possible extension is designing specialized encoders to extend scProTrans to more omics, including spatial transcriptome, spatial proteome, etc.

Conclusions

In summary, scProTrans provides an interpretable and sequence-aware deep learning framework for cross-omics translation at single-cell resolution. By integrating prior sequence knowledge with multi-omics data, scProTrans enables accurate and robust protein abundance prediction from scRNA-seq data, including zero-shot translation for proteins that can not directly measured. Extensive evaluations across diverse datasets demonstrate its superiority over existing methods and its utility in downstream analyses such as cell clustering, subtype identification, biomarker discovery, and perturbation-response analysis. Furthermore, its modular design allows extension to more complex multi-omics settings, highlighting its potential as a versatile platform for future single-cell multi-modal studies.

Methods

scProTrans consists of five modules: gene encoder, protein encoder, cell encoder, cross-omics translation module and loss module. Next, we will describe each module in detail.

Gene encoder

From the NCBI database, we collected 58,766 human gene sequences. To address the issue of large variations in gene sequence lengths, we quantified the proportion of genes with different length intervals (Additional file 2: Fig. S14). The results show that the majority of genes are between 1,000 and 10,000 pb in length, and 60% of them are shorter than 10,000 pb. We cropped all gene sequences to 10,000 bp to cover most gene information and reduce memory usage. In short sequences, 'N' is used to pad their embedding, so its value is zero. In genomic studies, DNA2vec [18] has been widely used as a pre-trained model to generate k-mer representations [52, 53]. Using sliding windows, we divided the gene sequences into 100-length fragments and then into k-mers (k=8 in this paper). As an example, "TGATATCCTG" is divided into "TGATATCC", "GATATCCT", and "ATATCCTG". To obtain 100-length vectors for each fragment, we converted k-mers into 100-length vectors through dna2vec's pre-trained dictionary. Each gene embedding consists of the vectors of all segments spliced in order. The embeddings of all genes were generated once and stored in the gene encoder module to avoid duplication of coding and improve translation efficiency.

Protein encoder

From the UniProt database, we compiled all known sequences for human proteins, totaling 20,412 items. Protein sequences are increasingly translated into numerical representations with pre-trained language models, instead of conventional one-hot coding methods [20, 54, 55]. ProtT5 [20] is a transformer-based protein language model based on 2.1 million proteins and 393 billion amino acids. The whole sequence is treated as one sentence, and amino acids are treated as tokens, then the final layer of the transformer's attention stack is accessed to generate a 1,024-length context-aware embedding. In the case that the length of a protein sequence is n, the amino acid embeddings are stacked into a matrix with $n \times 1024$, and finally averaged to obtain the 1,024-length protein embedding. This embedding has proved useful in facilitating proteome-related tasks such as protein function prediction, binding site discrimination, protein-peptide interaction site identification, etc. [56, 57]. All ProtT5-encoded embeddings of human proteins are assembled by the protein encoder module. By incorporating this prior knowledge, scProTrans is able to realize translations of all human proteins, even those that weren't included in the training set.

Cell encoder

As part of the cell encoder, raw transcriptome profiles are transformed into normalized profiles, and cell aspects are encoded to develop cell embeddings. Transcriptome profile raw count matrix was reduced by removing genes expressed in fewer than 10 cells, reducing data noise and focusing on widely expressed genes. In order to achieve comparable gene expression, the total number of cells was scaled to 10,000, eliminating the bias in expression between cells caused by the difference in sequencing depth. Following this, each count is transformed into a logarithmic number by adding 1. To create a normalized profile, we filtered the top 5000 highly variable genes.

Using normalized transcriptome profiles, we generated cell embeddings using the scVI algorithm proposed by Lopez et al. [58]. To achieve batch correction and noise reduction, scVI assumes gene expression follows a zero-inflated negative binomial distribution based on deep neural networks and variational self-encoders. A quintessential tool for processing single-cell RNA sequencing data, where the low-dimensional cell embedding facilitates a variety of basic analysis tasks [59, 60], and also serves as an input for the cross-omics translation module. As a default, 100 is the dimension of the cell embedding.

Cross-omics translation module

A cross-omics translation module requires four inputs: gene embedding from a gene encoder, protein embedding from a protein encoder, and cell embedding and normalized profile from a cell encoder. As opposed than gene embeddings derived directly from gene encoders, we extracted only a portion corresponding to genes included in the normalized profile. This is denoted by $G \in \mathbb{R}^{n_g \times d_g}$, where n_g is the number of genes and d_g is the vector length of each gene embedding. Here, g_{ij} represents the element in the i_{th} row and j_{th} column of the matrix G ($i = 1, \ldots, n_g$, $j = 1, \ldots, d_g$). We employed similar procedures to obtain protein embedding. It is denoted by $P \in \mathbb{R}^{n_p \times d_p}$, where n_p is the number

of proteins and d_p is the vector length of each protein embedding. Here, p_{ij} represents the element in the i_{th} row and j_{th} column of the matrix P ($i = 1, \ldots, n_p$, $j = 1, \ldots, d_p$). Cell embedding is represented as $C \in \mathbb{R}^{n_c \times d_c}$, where n_c is the number of cells and d_c is the vector length of each cell embedding. Here, c_{ij} represents the element in the i_{th} row and j_{th} column of the matrix C ($i = 1, \ldots, n_c$, $j = 1, \ldots, d_c$). Then, $X \in \mathbb{R}^{n_c \times n_g}$ denotes the normalized transcriptome profile.

Gene embedding and protein embedding are identical in different cells. Cell heterogeneity was captured by including cell-specific embeddings. Using replication, we extended the 2-dimensional matrix G to the 3-dimensional matrix $\widetilde{G} \in \mathbb{R}^{n_c \times n_g \times d_g}$, where n_c is the number of cells. For an element $\widetilde{g_{kij}} \in \widetilde{G}$, its correspondence with G can be expressed as:

$$\widetilde{g}_{kij} = g_{ij} \tag{1}$$

where $\widetilde{g}_{kij}$ is essentially a repeated representation of g_{ij} across k. Similarly, P is extended to a 3-dimensional matrix $\widetilde{P} \in \mathbb{R}^{n_c \times n_p \times d_p}$ with an element denoted $\widetilde{p_{kij}}$. Moreover, C is extended to $\widetilde{C^g} \in \mathbb{R}^{n_c \times n_g \times d_c}$. For an element $\widetilde{c^g_{ikj}} \in \widetilde{C^g}$, its correspondence with C can be expressed as:

$$\widetilde{c^p_{ikj}} = c_{ij} \tag{2}$$

We concatenated $\widetilde{G}$ and $\widetilde{C^g}$ along the third dimension to obtain the gene-cell combination matrix $M^{gc} \in \mathbb{R}^{n_c \times n_g \times d_{gc}}$, where $d_{gc} = d_g + d_c$. The element $m^{gc}_{ijk} \in M^{gc}$ can be calculated by:

$$m^{gc}_{ijk} = \begin{cases} \widetilde{g_{ijk}}, & k \le d_g \\ \widetilde{c^g_{ij(k-d_g)}}, & k > d_g \end{cases} \tag{3}$$

Next, we concatenated $\widetilde{P}$ and $\widetilde{C^p}$ along the third dimension to obtain the protein-cell combination matrix $M^{pc} \in \mathbb{R}^{n_c \times n_p \times d_{pc}}$, where $d_{pc} = d_p + d_c$. The element $m^{pc}_{ijk} \in M^{pc}$ can be calculated by:

$$m^{pc}_{ijk} = \begin{cases} \widetilde{p_{ijk}}, & k \le d_p \\ \widetilde{c^p_{ij(k-d_p)}}, & k > d_p \end{cases} \tag{4}$$

Gene-cell combination matrix serves as key, protein-cell combination matrix serves as query, and normalized transcriptome profile serves as value. We used the multi-head cross-attention mechanism to predict protein expression values in each cell. Query, key and value of the cell i correspond to M^{pc}_i, M^{gc}_i and X_i, which are briefly denoted by Q, K and V below. We used a linear layer to obtain Q' and K' by dimensionality reduction of Q and K, which are then split to belong to different heads, denoted as $Q' = [Q_1, Q_2, \ldots, Q_{head}]$ and $K' = [K_1, K_2, \ldots, K_{head}]$, where $head$ represents the number of heads in the multi-head attention mechanism and that is set to 10 in scProTrans. The attention score is calculated as follows:

$$Attention(Q, K, V) = soft\max\left(\frac{QK^T}{\sqrt{d}}\right)V \tag{5}$$

$$head_i = Attention(Q_i, K_i, V) \tag{6}$$

$$MultiHead\left(Q', K', V\right) = Concat(head_1, head_2, \ldots, head_h) \tag{7}$$

Through a two-layer linear neural network, the output of the multi-head attention mechanism is mapped to a vector $\widehat{Y_i} = [\widehat{y_{i1}}, \widehat{y_{i2}}, \ldots, \widehat{y_{in_p}}]$, which corresponds to the expression levels of each protein in the cell i.

Loss module

Loss module processes raw proteome profiles to generate normalized profiles, and calculates loss to optimize model parameters. After scaling the raw count matrix to 10,000, we log-transformed each count after adding 1 and obtained the normalized proteome profile. Here, $Y \in \mathbb{R}^{n_c \times n_p}$ denotes the normalized proteome profile and y_{ij} represents the expression value of protein j in cell i. In order to compute loss, we use the normalized proteome profile as well as a predicted proteome profile developed from the cross-omics translation module. The loss function is:

$$Loss = \sum_{i=1}^{n_c} \left(\frac{1}{n_p} \sum_{j=1}^{n_p} \left(\widehat{y_{ij}} - y_{ij}\right)^2 \right) \tag{8}$$

where $\widehat{y_{ij}}$ denotes the predicted expression value of protein j in cell i, and y_{ij} denotes the ground truth. As a default, the epoch used during training is 200. To avoid overfitting, our model stops training after 50 consecutive epochs if loss does not decrease.

Translation process

scProTrans can translate paired proteome profiles based on newly sequenced transcriptome profiles. During the translation process, scProTrans filters genes from transcriptome data that are identical to the training set to match the well-trained gene-protein associations. By default, scProTrans utilizes protein embeddings from the training set to translate matching protein expression profiles. In zero-shot translation mode, the protein encoder relies on a priori sequence knowledge to generate new protein embeddings according to the specified proteins, enabling the prediction of unavailable proteins in the training set.

Evaluation metrics

In this paper, Mean Square Error (MSE), Mean Absolute Error (MAE) and Correlation are utilized to evaluate different tools for the proteome translation task. The formulas are:

$$MSE = \frac{1}{n_p} \sum_{j=1}^{n_p} \left(\frac{1}{n_c} \sum_{i=1}^{n_c} \left(\widehat{y_{ij}} - y_{ij}\right)^2 \right) \tag{9}$$

$$MAE = \frac{1}{n_p}\sum_{j=1}^{n_p}\left(\frac{1}{n_c}\sum_{i=1}^{n_c}\left|\widehat{y_{ij}} - y_{ij}\right|\right) \tag{10}$$

$$Correlation = \frac{1}{n_p}\sum_{j=1}^{n_p}\left(\frac{\sum_{i=1}^{n_c}\left(\widehat{y_{ij}} \times y_{ij}\right)}{\sqrt{\sum_{i=1}^{n_c}\widehat{y_{ij}}^2} \times \sqrt{\sum_{i=1}^{n_c} y_{ij}^2}}\right) \tag{11}$$

where n_c is the number of cells and n_p is the number of proteins.

In comparing cell clustering based on transcriptome, proteome, and translation data, we employed Principal Component Analysis (PCA) algorithm to reduce the dimensionality of expression profiles to 50, and then constructed neighborhood graphs and conducted Leiden clustering [61]. A series of metrics were used to evaluate the clustering results, including Rand Index (RI), Adjusted Rand Index (ARI), Adjusted Mutual Information (AMI), Normalized Mutual Information (NMI), Homogeneity (HOM), and V-measure (V), all of which are implemented by scikit-learn [62]. By using these metrics, we can assess the clustering outcomes from multiple perspectives, including agreement, purity, information content, and compactness. As each metric addresses a different aspect of clustering quality, a thorough analysis and comparison can be made across different algorithms or datasets.

Single-cell gene-protein association

We adopted a multi-head attention mechanism for the cross-omics translation module. Different cells correspond to the same gene embeddings and protein embeddings, but different cell embeddings. Since Query and Key are combinations of protein embedding with cell embedding and gene embedding with cell embedding, respectively, they are cell-specific, and the attention matrix reflects this. As the cell-specific gene-protein association matrix, we summed the attention matrices corresponding to different heads. In addition, we calculated the cell type specific gene-protein association by summing the cell-specific gene-protein association matrices belonging to the same cell type. Those genes and proteins with high values are more likely to play an important role in the translation process. In cell-specific matrices, cell-type patterns are reflected as well as differences among cells in a few genes and proteins. Figure 4 validates the possibility that marker genes and proteins are associated with high values in the cell type gene-protein association.

When identifying marker genes, CellMarker2.0 [29] records serve as the gold standard. As thresholds, we use the mean log foldchange value and attention score of the top 50 differentially expressed genes. When only considering foldchange, genes with values higher than the foldchange threshold are considered markers. When both foldchange and attention score are taken into account, genes with both values above their thresholds are designated as marker genes. The formula for calculating precision is as follows:

$$Precision = \frac{TP}{TP + FP} \tag{12}$$

where *TP* indicates the number of predicted marker genes in the gold standard, and *FP* indicates the number of predicted marker genes not in the gold standard.

Zero-shot translation mechanism

In the zero-shot translation mechanism, scProTrans translates protein profiles unavailable from multi-omics sequencing technologies. The sequence information is tightly related to expression abundance, and many methods have been developed to predict expression levels based on gene and protein sequences [63–65]. A series of paired transcriptome-proteome datasets are used in scProTrans to train its linkage of gene sequences, protein sequences, and expression patterns. Protein encoder and gene encoder construct gene embeddings and protein embeddings based on prior sequence knowledge. To link sequences to expression profiles, the cross-omics translation module combines sequence embeddings with cell embeddings. While some proteins are difficult to capture by multi-omics sequencing technologies, proteins with similar sequences are likely to be included in reference datasets.

scProTrans translates the expression profiles of unseen proteins by mapping their sequence-derived embeddings to those of seen proteins with similar sequences, thereby inferring their expression patterns. The protein encoder in scProTrans is pretrained on the full set of human protein sequences, generating fixed-length embeddings for each protein. These embeddings are derived from prior knowledge and independent from multi-omics datasets. During model training, the cross-omics translation module learns the mapping from transcriptome to proteome via an attention mechanism over cell embeddings C, gene embeddings G and protein embeddings corresponding to seen proteins (which are included in training set) P_{seen}, which can be expressed as:

$$\widehat{Y} = F_{train}(C,\ G,\ P_{seen}) \tag{13}$$

where $\widehat{Y}$ denotes the predicted proteome profile, F denotes the cross-omics translation module.

In the zero-shot translation phase, protein encoder extracts the embeddings of unseen proteins (which are unavailable in training set and also prediction targets) P_{unseen}. The cross-omics translation module takes the cell embeddings C, gene embeddings G and protein embeddings P_{unseen} as input. The expression of the unseen proteins is then predicted as:

$$\widehat{Y} = F_{zero-shot}(C,\ G,\ P_{unseen}) \tag{14}$$

This design allows the model to leverage sequence and expression similarities between seen and unseen proteins, enabling effective prediction of proteins not present in the training set and achieving zero-shot generalization. As illustrated in Fig. 5B, C, and D, the CD324 protein—absent from the training set—serves as a compelling case study. Here, scProTrans utilizes the CD325 protein as a reference to demonstrate the stepwise translation process, from sequence extraction to embedding generation and ultimately to the predicted expression profile.

Extending scProTrans to epigenome

By adding appropriate encoders to scProTrans, it's easy to migrate it to other omics translation tasks because of its modular structure and cross-omics translation core. In epigenome, scATAC-seq data serve as a characterisation of chromatin accessibility as well as a tool for understanding regulatory elements and mechanisms [66, 67]. A peak encoder is used to introduce prior sequence knowledge into the scATAC-seq dataset. Using their locations, we mapped peaks to human reference genomes to obtain sequence fragments. For short sequences, we filled 'N' to make their embedding values 0 by cropping all sequences to 2,000 bp. As a result, dna2vec algorithm encodes each sequence into a 2000-long peak embedding.

Additionally, the cell encoder for epigenome generates normalized epigenome profiles and cell embeddings from scATAC-seq data. Let $X^{n_c \times n_p}$ denote the raw profile containing n_p peaks and n_c cells, where x_{ij} represents the value of the peak j in cell i. In order to analyze the raw count matrix, we performed the term frequency-inverse document frequency (TF-IDF) transformation widely used for scATAC-seq data analysis [68, 69]. Here is the formula:

$$x'_{ij} = \frac{x_{ij}}{\sum_{k=1}^{n_p} x_{ik}} \times \log\left(1 + \frac{n_c}{\sum_{k=1}^{n_c} x_{kj}}\right) \tag{15}$$

This procedure normalizes the depth of sequencing and assigns a higher value to rare peaks. Then, we extracted the top 5,000 highly variable peaks in order to produce a normalized epigenome profile. Finally, a normalized epigenome profile is used to create cell embeddings by scVI. In the epigenome-to-transcriptome translation task, the encoders in scProTrans include a peak encoder, a cell encoder for epigenome, and a gene encoder. The peak embedding combined with cell embedding represents as Key, the gene embedding combined with cell embedding represents as Query, and the normalized epigenome profile represents as Value, which comprises the input to the cross-omics translation module. In the epigenome-to-proteome translation task, the encoders in scProTrans include a peak encoder, a cell encoder for epigenome, and a protein encoder. The peak embedding combined with cell embedding represents as Key, the protein embedding combined with cell embedding represents as Query, and the normalized epigenome profile represents as Value, which comprises the input to the cross-omics translation module. An overview of the modeling structure can be found in Additional file 2: Fig. S15.

Datasets

To comprehensively evaluate the performance of our proposed method, we curated a diverse collection of single-cell omics datasets, spanning multiple experimental techniques and biological contexts. The dataset cohort encompasses both single-cell multi-omics and single-cell RNA-seq data, providing a rich resource for rigorous validation.

Single-cell multi-omics datasets

GSE194122 [70] was generated to support the Multimodal Single-Cell Data Integration Challenge at NeurIPS 2021. By CITE-seq, human bone marrow mononuclear cells (BMMCs) were collected from 9 healthy donors and four sites. In total, 45 cell types and 12 batches were annotated. As a result of removing the holdout cells, the dataset consisted of 81,241 cells and 134 proteins.

GSE100866 [9] was derived from human cord blood mononuclear cells (CBMCs) by CITE-seq. There are 8,617 cells and 13 proteins, respectively.

GSE164378 [23] was derived from Hao et al., who performed CITE-seq on peripheral blood mononuclear cells (PBMCs) from 8 volunteers enrolled in an HIV vaccine trial. Cells were collected after patients received a VSV-vectored HIV vaccine at 0, 3, and 7 days. This dataset contains 161,764 cells and 228 proteins, with 3 hierarchies of cell type annotation and 2 batches.

GSE128639 [71] was derived from Stuart et al., who performed CITE-seq on human bone marrow cells. It contains 33,454 cells and 25 proteins.

GSM4732113 and GSM4732115 are CITE-seq datasets from PBMCs [72]. Cells in GSM4732115 were stimulated with anti-CD3/CD28 and IL-2 for 16 h. Cells in GSM4732113 were cultured without stimulation. GSM4732113 has 5,527 cells and 227 proteins, whereas GSM4732115 has 4,115 cells and 227 proteins.

GSE200417_CITE and GSE200417_DOGMA represent the paired CITE-seq and DOGMA-seq datasets, respectively, from peripheral blood T cells [47]. GSE200417_CITE includes 14,821 cells and 116 proteins, while GSE200417_DOGMA includes 14,172 cells and 163 proteins.

GSM4949911, GSM5123951, GSM5123952, GSM5123953, GSM5123954, and GSM5123955 were derived from Swanson et al. [48], and belong to human PBMCs. Among them, GSM5123955 is CITE-seq dataset, which contains 7657 cells and 46 proteins. The others are TEA-seq datasets that provide matched epigenome, transcriptome, and proteome data. The number of cells involved in GSM4949911, GSM5123951, GSM5123952, and GSM5123953 are 8213, 7966, 8512, 8496 and 9240, respectively.

GSM6578065 is a spatial CITE-seq dataset from human skin [49]. It contains 1,691 spatial spots, 283 proteins, and 15,486 genes.

For each dataset, the above number of cells is the result of aligning different omics data. Specifically, cells detected exclusively in a single omics modality were excluded to ensure only cells with corresponding multi-omics measurements were retained. Comprehensive details regarding dataset composition, including cell types, batches, and references, are provided in Additional file 1: Table S1.

Single-cell RNA-seq datasets

GSM2560248 and GSM2560249 are scRNA-seq datasets of PBMCs from 8 lupus patients [37]. GSM2560248 includes 14,619 cells with a control condition while GSM2560249 includes 14,446 cells that were stimulated with IFN-β.

Implementation details and hyperparameter settings

In this section, we provide a detailed description of the implementation of scProTrans, including its hyperparameter settings and software configurations. Specifically, the gene encoder transforms each gene into a 10,000-dimensional vector as the gene embedding. The protein encoder encodes each protein sequence into a 1,024-dimensional protein embedding. The cell encoder leverages scVI to reduce each cell to a 100-dimensional latent embedding.

The scVI model uses default parameters. Specifically, the latent dimension is set to 100, with other parameters and priors kept at their default values, including a hidden layer dimension of 128, one hidden layer, and a dropout rate of 0.1. We also use the default library size normalization in scVI, i.e., using the observed library size for RNA as the scaling factor in the mean of the conditional distribution. All prior distributions in scVI are kept at default, including modeling gene expression with a zero-inflated negative binomial (ZINB) distribution. These settings ensure that scVI processes transcriptomic data under standard and stable conditions.

The Cross-omics Translation Module consists of a linear mapping layer, a multi-head attention module, and a multi-layer feedforward prediction network. In detail, the linear mapping layer unifies inputs of different dimensions into a fixed hidden dimension of 500 to facilitate cross-omics attention mechanism. The multi-head attention module includes 10 heads, each with a dimensionality of 50. Following the attention output, the model employs two successive linear layers for feature fusion and non-linear mapping, each followed by a ReLU activation, with hidden dimensions of 10 and 1, respectively.

During training, scProTrans uses the Adam optimizer. Key hyperparameters are set as follows: a batch size of 32, 200 training epochs, a learning rate of 0.001, an early stopping patience of 50, and a random seed of 0 to ensure reproducibility. For the t-SNE visualization analysis, we adopt the TSNE module in scikit-learn [62] with the following settings: n_components $= 2$, init =“pca”, and random_state $= 0$. In the “Proteomic perturbation response analysis Via scProTrans” section, EdgeR [73] and a threshold of $-\log_{10}(P - Value) > 10$ are used to identify differentially expressed proteins. The model is implemented in PyTorch 2.0, running on Python 3.9 with CUDA 11.7 on an NVIDIA A100 GPU for training and evaluation.

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13059-026-04070-6.

Additional file 1: Supplementary tables. This file contains supplementary tables S1-S6.

Additional file 2: Supplementary notes and figures. This file contains supplementary notes S1-S6 and supplementary figures S1-S15.

Acknowledgements
Not applicable.

Peer review information
Shila Ghazanfar and Tim Sands were the primary editors of this article and managed its editorial process and peer review in collaboration with the rest of the editorial team. The peer-review history is available in the online version of this article.

Authors' contributions
M.Z. contributed to the methodology, result analysis and writing. J.L. contributed to the datasets collection and model implementation. Y.J., X.T. and J.Y. performed the comparison experiments. F.G. and J.T. supervised the study. All authors read and approved the final manuscript.

Funding
This work was supported by the National Natural Science Foundation of China (Grants No. U24A20257 to J.T., No. 62322215 and No. 62532017 to F.G.), Shenzhen Science and Technology Program (Grants No. JCYJ20241202130212016 and No. KQTD20200820113106007 to J.T., and No. JCYJ20230807140709020 to M.Z.), Natural Science Foundation of Hunan Province (Grants No. 2026JJ30018 to F.G.). This study was also supported in part by the High-Performance Computing Center of Central South University.

Data availability
The scProTrans is available on GitHub (https://github.com/MengyuanZhaoo/ProTrans) [74] and Zenodo (https://doi.org/10.5281/zenodo.19216928) [75] under MIT license. The pretrained gene sequence embeddings are available at this link. The pretrained protein sequence embeddings are available at this link. The pretrained embeddings are also available on Zenodo (https://doi.org/10.5281/zenodo.19217421 [76]. All the datasets used in this paper were previously published and freely available. They are in the GEO database (https://www.ncbi.nlm.nih.gov/geo/) under accession codes corresponding to GSE199424, GSE100866, GSE164378, GSE128639, GSE156473, GSE200417, GSE158013, GSE213264 and GSE96583.

Declarations

Ethics approval and consent to participate
Not applicable.

Competing interests
The authors declare no competing interests.

Received: 20 August 2025 Accepted: 2 April 2026
Published online: 14 April 2026

References

1. Argelaguet R, Cuomo ASE, Stegle O, Marioni JC. Computational principles and challenges in single-cell data integration. Nat Biotechnol. 2021;39(10):1202–15. https://doi.org/10.1038/s41587-021-00895-7.
2. Eisele AS, Tarbier M, Dormann AA, Pelechano V, Suter DM. Gene-expression memory-based prediction of cell lineages from scRNA-seq datasets. Nat Commun. 2024;15(1). https://doi.org/10.1038/s41467-024-47158-y.
3. Zhao M, He W, Tang J, Zou Q, Guo F. A hybrid deep learning framework for gene regulatory network inference from single-cell transcriptomic data. Brief Bioinform. 2022;23(2). https://doi.org/10.1093/bib/bbab568.
4. Labib M, Kelley SO. Single-cell analysis targeting the proteome. Nat Rev Chem. 2020;4(3):143–58.
5. Brunner A, Thielert M, Vasilopoulou C, Ammar C, Coscia F, Mund A, et al. Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. Mol Syst Biol. 2022;18(3):e10798. https://doi.org/10.15252/msb.202110798.
6. Slavov N. Unpicking the proteome in single cells. Science. 2020;367(6477):512–3. https://doi.org/10.1126/science.aaz6695.
7. Schoof EM, Furtwängler B, Üresin N, Rapin N, Savickas S, Gentil C, et al. Quantitative single-cell proteomics as a tool to characterize cellular hierarchies. Nat Commun. 2021;12(1):3341.
8. Furnari FB, Cloughesy TF, Cavenee WK, Mischel PS. Heterogeneity of epidermal growth factor receptor signalling networks in glioblastoma. Nat Rev Cancer. 2015;15(5):302–10.
9. Stoeckius M, Hafemeister C, Stephenson W, Houck-Loomis B, Chattopadhyay PK, Swerdlow H, et al. Simultaneous epitope and transcriptome measurement in single cells. Nat Methods. 2017;14(9):865. https://doi.org/10.1038/NMETH.4380.
10. Peterson VM, Zhang KX, Kumar N, Wong J, Li L, Wilson DC, et al. Multiplexed quantification of proteins and transcripts in single cells. Nat Biotechnol. 2017;35(10):936. https://doi.org/10.1038/nbt.3973.
11. Baysoy A, Bai Z, Satija R, Fan R. The technological landscape and applications of single-cell multi-omics. Nat Rev Mol Cell Biol. 2023;24(10):695–713. https://doi.org/10.1038/s41580-023-00615-w.
12. Zhou M, Zhang H, Bai Z, Mann-Krzisnik D, Wang F, Li Y. Single-cell multi-omics topic embedding reveals cell-type-specific and COVID-19 severity-related immune signatures. Cell Rep Methods. 2023;3(8). https://doi.org/10.1016/j.crmeth.2023.100563.
13. Du JH, Cai Z, Roeder K. Robust probabilistic modeling for single-cell multimodal mosaic integration and imputation via scVAEIT. Proc Natl Acad Sci USA. 2022;119(49). https://doi.org/10.1073/pnas.2214414119.
14. Gayoso A, Steier Z, Lopez R, Regier J, Nazor KL, Streets A, et al. Joint probabilistic modeling of single-cell multi-omic data with totalVI. Nat Methods. 2021;18(3):272. https://doi.org/10.1038/s41592-020-01050-x.
15. Cao Y, Zhao X, Tang S, Jiang Q, Li S, Li S, et al. scButterfly: a versatile single-cell cross-modality translation method via dual-aligned variational autoencoders. Nat Commun. 2024;15(1). https://doi.org/10.1038/s41467-024-47418-x.
16. Liu L, Li W, Wong KC, Yang F, Yao J. A pre-trained large language model for translating single-cell transcriptome to proteome. bioRxiv. 2023. https://doi.org/10.1101/2023.07.04.547619.
17. Sayers EW, Beck J, Bolton EE, Brister JR, Chan J, Comeau DC, et al. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2024;52(D1):D33–43. https://doi.org/10.1093/nar/gkad1044.
18. Ng P. dna2vec: Consistent vector representations of variable-length k-mers. arXiv. 2017. https://arxiv.org/abs/1701.06279.
19. Consortium TU. UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Res. 2024;gkae1010. https://doi.org/10.1093/nar/gkae1010.

20. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, et al. Prottrans: Toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell. 2021;44(10):7112–27.
21. Lakkis J, Schroeder A, Su K, Lee MYY, Bashore AC, Reilly MP, 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. 2022;4(11):940. https://doi.org/10.1038/s42256-022-00545-w.
22. Hao Y, Stuart T, Kowalski MHH, Choudhary S, Hoffman P, Hartman A, et al. Dictionary learning for integrative, multimodal and scalable single-cell analysis. Nat Biotechnol. 2024;42(2). https://doi.org/10.1038/s41587-023-01767-y.
23. Hao Y, Hao S, Andersen-Nissen E, Mauck WMIIIIII, Zheng S, Butler A, et al. Integrated analysis of multimodal single-cell data. Cell. 2021;184(13):3573. https://doi.org/10.1016/j.cell.2021.04.048.
24. Wolf FA, Angerer P, Theis FJ. SCANPY: large-scale single-cell gene expression data analysis. Genome Biol. 2018;19. https://doi.org/10.1186/s13059-017-1382-0.
25. Martrus G, Goebels H, Langeneckert AE, Kah J, Flomm F, Ziegler AE, et al. CD49a Expression Identifies a Subset of Intrahepatic Macrophages in Humans. Front Immunol. 2019;10. https://doi.org/10.3389/fimmu.2019.01247.
26. El-Bassiouni N, Messery L, Zayed R, Metwally O, Zahran M, Mahmoud O, et al. Tissue factor expression on blood monocytes in patients with hepatitis C virus-induced chronic liver disease. Comp Clin Pathol. 2014;23:1159–66. https://doi.org/10.1007/s00580-013-1757-x.
27. Ravkov EV, Williams ESCP, Elgort M, Barker AP, Planelles V, Spivak AM, et al. Reduced monocyte proportions and responsiveness in convalescent COVID-19 patients. Front Immunol. 2024;14. https://doi.org/10.3389/fimmu.2023.1329026.
28. Affandi AJ, Olesek K, Grabowska J, Twilhaar MKN, Rodriguez E, Saris A, et al. CD169 Defines Activated CD14^{+} Monocytes With Enhanced CD8^{+} T Cell Activation Capacity. Front Immunol. 2021;12. https://doi.org/10.3389/fimmu.2021.697840.
29. Hu C, Li T, Xu Y, Zhang X, Li F, Bai J, et al. CellMarker 2.0: an updated database of manually curated cell markers in human/mouse and web tools based on scRNA-seq data. Nucleic Acids Res. 2023;51(D1):D870–D876. https://doi.org/10.1093/nar/gkac947.
30. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, et al. Clustal W and clustal X version 2.0. Bioinformatics. 2007;23(21):2947–8. https://doi.org/10.1093/bioinformatics/btm404.
31. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. Jalview Version 2-a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009;25(9):1189–91. https://doi.org/10.1093/bioinformatics/btp033.
32. Lotfollahi M, Wolf FA, Theis FJ. scGen predicts single-cell perturbation responses. Nat Methods. 2019;16(8):715. https://doi.org/10.1038/s41592-019-0494-8.
33. Bunne C, Stark SG, Gut G, del Castillo JS, Levesque M, Lehmann KV, et al. Learning single-cell perturbation responses using neural optimal transport. Nat Methods. 2023;20(11). https://doi.org/10.1038/s41592-023-01969-x.
34. Anchang B, Davis KL, Fienberg HG, Williamson BD, Bendall SC, Karacosta LG, et al. DRUG-NEM: Optimizing drug combinations using single-cell perturbation response to account for intratumoral heterogeneity. Proc Natl Acad Sci USA. 2018;115(18):E4294–303. https://doi.org/10.1073/pnas.1711365115.
35. Schubert M, Klinger B, Kluenemann M, Sieber A, Uhlitz F, Sauer S, et al. Perturbation-response genes reveal signaling footprints in cancer gene expression. Nat Commun. 2018;9. https://doi.org/10.1038/s41467-017-02391-6.
36. Ji Y, Lotfollahi M, Wolf FA, Theis FJ. Machine learning for perturbational single-cell omics. Cell Syst. 2021;12(6):522–37. https://doi.org/10.1016/j.cels.2021.05.016.
37. Kang HM, Subramaniam M, Targ S, Nguyen M, Maliskova L, McCarthy E, et al. Multiplexed droplet single-cell RNA-sequencing using natural genetic variation. Nat Biotechnol. 2018;36(1):89. https://doi.org/10.1038/nbt.4042.
38. Li X, Wang K, Lyu Y, Pan H, Zhang J, Stambolian D, et al. Deep learning enables accurate clustering with batch effect removal in single-cell RNA-seq analysis. Nat Commun. 2020;11(1). https://doi.org/10.1038/s41467-020-15851-3.
39. Lotfollahi M, Rybakov S, Hrovatin K, Hediyeh-zadeh S, Talavera-Lopez C, Misharin AV, et al. Biologically informed deep learning to query gene programs in single-cell atlases. Nat Cell Biol. 2023;25(2):337. https://doi.org/10.1038/s41556-022-01072-x.
40. Li Y, Lee PY, Reeves WH. Monocyte and Macrophage Abnormalities in Systemic Lupus Erythematosus. Arch Immunol Ther Exp. 2010;58(5):355–64. https://doi.org/10.1007/s00005-010-0093-y.
41. Wang D, Chen K, Du WT, Han ZB, Ren H, Chi Y, et al. CD14+ monocytes promote the immunosuppressive effect of human umbilical cord matrix stem cells. Exp Cell Res. 2010;316(15):2414–23. https://doi.org/10.1016/j.yexcr.2010.04.018.
42. Williams AEG, Choi K, Chan AL, Lee YJ, Reeves WH, Bubb MR, et al. Sjogren's syndrome-associated microRNAs in CD14+ monocytes unveils targeted TGFβ signaling. Arthritis Res Ther. 2016;18. https://doi.org/10.1186/s13075-016-0987-0.
43. Zhong H, Bao W, Li X, Miller A, Seery C, Haq N, et al. CD16+ monocytes control T-cell subset development in immune thrombocytopenia. Blood. 2012;120(16):3326–35. https://doi.org/10.1182/blood-2012-06-434605.
44. Iacobas D, Wen J, Iacobas S, Schwartz N, Putterman C. Remodeling of Neurotransmission, Chemokine, and PI3K-AKT Signaling Genomic Fabrics in Neuropsychiatric Systemic Lupus Erythematosus. Genes. 2021;12(2). https://doi.org/10.3390/genes12020251.
45. Stylianou K, Petrakis I, Mavroeidi V, Stratakis S, Vardaki E, Perakis K, et al. The PI3K/Akt/mTOR pathway is activated in murine lupus nephritis and downregulated by rapamycin. Nephrol Dial Transplant. 2011;26(2):498–508. https://doi.org/10.1093/ndt/gfq496.
46. Zhao C, Gu Y, Chen L, Su X. Upregulation of FoxO3a expression through PI3K/Akt pathway attenuates the progression of lupus nephritis in MRL/lpr mice. Int Immunopharmacol. 2020;89(A). https://doi.org/10.1016/j.intimp.2020.107027.
47. Xu Z, Heidrich-O'Hare E, Chen W, Duerr RH. Comprehensive benchmarking of CITE-seq versus DOGMA-seq single cell multimodal omics. Genome Biol. 2022;23(1). https://doi.org/10.1186/s13059-022-02698-8.
48. Swanson E, Lord C, Reading J, Heubeck AT, Genge PC, Thomson Z, et al. Simultaneous trimodal single-cell measurement of transcripts, epitopes, and chromatin accessibility using TEA-seq. eLife. 2021;10. https://doi.org/10.7554/eLife.63632.

49. Liu Y, DiStasio M, Su G, Asashima H, Enninful A, Qin X, et al. High-plex protein and whole transcriptome co-mapping at cellular resolution with spatial CITE-seq. Nat Biotechnol. 2023;41(10):1405. https://doi.org/10.1038/s41587-023-01676-0.
50. Benoit BM, Jariwala N, O'Connor G, Oetjen LK, Whelan TM, Werth A, et al. CD164 identifies CD4+ T cells highly expressing genes associated with malignancy in Sezary syndrome: the Sezary signature genes, FCRL3, Tox, and miR-214. Arch Dermatol Res. 2017;309(1):11–9. https://doi.org/10.1007/s00403-016-1698-8.
51. Sachdev R, George TI, Schwartz EJ, Sundram UN. Discordant Immunophenotypic Profiles of Adhesion Molecules and Cytokines in Acute Myeloid Leukemia Involving Bone Marrow and Skin. Am J Clin Pathol. 2012;138(2):290–9. https://doi.org/10.1309/AJCP34YERPZSCYKQ.
52. He W, Zhou H, Zuo Y, Bai Y, Guo F. MuSE: A deep learning model based on multi-feature fusion for super-enhancer prediction. Comput Biol Chem. 2024;113. https://doi.org/10.1016/j.compbiolchem.2024.108282.
53. Luo H, Li Y, Liu H, Ding P, Yu Y, Luo L. SENet: A deep learning framework for discriminating super- and typical enhancers by sequence information. Comput Biol Chem. 2023;105. https://doi.org/10.1016/j.compbiolchem.2023.107905.
54. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023;379(6637):1123–30. https://doi.org/10.1126/science.ade2574.
55. Brandes N, Ofer D, Peleg Y, Rappoport N, Linial M. ProteinBERT: a universal deep-learning model of protein sequence and function. Bioinformatics. 2022;38(8):2102–10. https://doi.org/10.1093/bioinformatics/btac020.
56. Fang Y, Jiang Y, Wei L, Ma Q, Ren Z, Yuan Q, et al. DeepProSite: structure-aware protein binding site prediction using ESMFold and pretrained language model. Bioinformatics. 2023;39(12). https://doi.org/10.1093/bioinformatics/btad718.
57. Le VT, Zhan ZJ, Vu TTP, Malik MS, Ou YY. ProtTrans and multi-window scanning convolutional neural networks for the prediction of protein-peptide interaction sites. J Mol Graph Model. 2024;130. https://doi.org/10.1016/j.jmgm.2024.108777.
58. Lopez R, Regier J, Cole MB, Jordan MI, Yosef N. Deep generative modeling for single-cell transcriptomics. Nat Methods. 2018;15(12):1053. https://doi.org/10.1038/s41592-018-0229-2.
59. Wang J, Zou Q, Lin C. A comparison of deep learning-based pre-processing and clustering approaches for single-cell RNA sequencing data. Brief Bioinform. 2022;23(1). https://doi.org/10.1093/bib/bbab345.
60. Brendel M, Su C, Bai Z, Zhang H, Elemento O, Wang F. Application of Deep Learning on Single-cell RNA Sequencing Data Analysis: A Review. Genomics Proteomics Bioinforma. 2022;20(5, SI):814–35. https://doi.org/10.1016/j.gpb.2022.11.011.
61. Traag VA, Waltman L, van Eck NJ. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep. 2019;9. https://doi.org/10.1038/s41598-019-41695-z.
62. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, et al. Scikit-learn: Machine Learning in Python. J Mach Learn Res. 2011;12:2825–30.
63. Stefanini M, Lovino M, Cucchiara R, Ficarra E. Predicting gene and protein expression levels from DNA and protein sequences with Perceiver. Comput Methods Programs Biomed. 2023;234:107504. https://doi.org/10.1016/j.cmpb.2023.107504.
64. Agarwal V, Shendure J. Predicting mRNA Abundance Directly from Genomic Sequence Using Deep Convolutional Neural Networks. Cell Rep. 2020;31(7). https://doi.org/10.1016/j.celrep.2020.107663.
65. Avsec Z, Agarwal V, Visentin D, Ledsam JR, Grabska-Barwinska A, Taylor KR, et al. Effective gene expression prediction from sequence by integrating long-range interactions. Nat Methods. 2021;18(10):1196. https://doi.org/10.1038/s41592-021-01252-x.
66. Malekpour SA, Haghverdi L, Sadeghi M. Single-cell multi-omics analysis identifies context-specific gene regulatory gates and mechanisms. Brief Bioinform. 2024;25(3). https://doi.org/10.1093/bib/bbae180.
67. Cao ZJ, Gao G. Multi-omics single-cell data integration and regulatory inference with graph-linked embedding. Nat Biotechnol. 2022;40(10):1458. https://doi.org/10.1038/s41587-022-01284-4.
68. Chen X, Chen S, Song S, Gao Z, Hou L, Zhang X, et al. Cell type annotation of single-cell chromatin accessibility data via supervised Bayesian embedding. Nat Mach Intell. 2022;4(2):116–26. https://doi.org/10.1038/s42256-021-00432-w.
69. Chen S, Wang R, Long W, Jiang R. ASTER: accurately estimating the number of cell types in single-cell chromatin accessibility data. Bioinformatics. 2023;39(1). https://doi.org/10.1093/bioinformatics/btac842.
70. Luecken M, Burkhardt D, Cannoodt R, Lance C, Agrawal A, Aliee H, et al. A sandbox for prediction and integration of DNA, RNA, and proteins in single cells. In: Vanschoren J, Yeung S, editors. Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks. vol. 1. Red Hook: Curran Associates, Inc.; 2021.
71. Stuart T, Butler A, Hoffman P, Hafemeister C, Papalexi E, Mauck WM III, et al. Comprehensive Integration of Single-Cell Data. Cell. 2019;177(7):1888. https://doi.org/10.1016/j.cell.2019.05.031.
72. Mimitou EP, Lareau CA, Chen KY, Zorzetto-Fernandes AL, Hao Y, Takeshima Y, et al. Scalable, multimodal profiling of chromatin accessibility, gene expression and protein levels in single cells. Nat Biotechnol. 2021;39(10):1246. https://doi.org/10.1038/s41587-021-00927-2.
73. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40. https://doi.org/10.1093/bioinformatics/btp616.
74. Zhao M, Li J, Jiang Y, Yan J, Tang X, Liang C, et al. A sequence knowledge-guided deep learning method for single-cell multi-omics translation. GitHub. 2026. https://github.com/MengyuanZhaoo/ProTrans. Accessed 10 Apr 2026.
75. Zhao M, Li J, Jiang Y, Yan J, Tang X, Liang C, et al. A sequence knowledge-guided deep learning method for single-cell multi-omics translation. Zenodo. 2026. https://doi.org/10.5281/zenodo.19216928.
76. Zhao M, Li J, Jiang Y, Yan J, Tang X, Liang C, et al. A sequence knowledge-guided deep learning method for single-cell multi-omics translation. Zenodo. 2026. https://doi.org/10.5281/zenodo.19217421.

Publisher's Note

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