Surface protein imputation from single cell transcriptomes by deep neural networks

Zilu Zhou[1,2], Chengzhong Ye[3], Jingshu Wang[2], Nancy R. Zhang[2*]

1) Graduate Group in Genomics and Computational Biology, University of Pennsylvania, Philadelphia, PA
2) Department of Statistics, University of Pennsylvania, Philadelphia, PA
3) School of Medicine, Tsinghua University, Beijing, China
* Correspondence:

Nancy R. Zhang
nzh@wharton.upenn.edu
(215) 898-8007
Department of Statistics
The Wharton School
University of Pennsylvania

While single cell RNA sequencing (scRNA-seq) is invaluable for studying cell populations, cell-surface proteins are often integral markers of cellular function and serve as primary targets for therapeutic intervention. Here we propose a transfer learning framework, single cell Transcriptome to Protein prediction with deep neural network (cTP-net), to impute surface protein abundances from scRNA-seq data by learning from existing single-cell multi-omic resources.

Recent technological advances allow the simultaneous profiling, across many cells in parallel, of multiple omics features in the same cell [1-5]. In particular, high throughput quantification of the transcriptome and a selected panel of cell surface proteins in the same cell is now feasible through the REAP-seq and CITE-seq protocols [2, 3]. Yet, due to technological barriers and cost considerations, most single cell studies, including Human Cell Atlas project [6], quantify the transcriptome only and do not have cell-matched measurements of relevant surface proteins, which can serve as integral markers of cellular function and primary targets for therapeutic intervention [7, 8]. Often, which proteins are relevant become apparent only after exploration by

scRNA-seq. This motivates our inquiry of whether protein abundances in individual cells can be accurately imputed by the cell's transcriptome. We propose cTP-net, a transfer learning approach based on deep neural networks, that imputes surface protein abundances for scRNA-seq data. cTP-net relies, for model training, on accumulating public data of cells with paired transcriptome and surface protein measurements.

Studies based on both CITE-seq and REAP-seq have shown that the relative abundance of most surface proteins, at the level of individual cells, is only weakly correlated with the relative abundance of the RNA of its corresponding gene [2, 3, 9]. This is due to technical factors such as RNA and protein measurement error [10], as well as inherent stochasticity in RNA processing, translation and protein transport [11-15]. To accurately impute surface protein abundance from scRNA-seq data, cTP-net employs two steps: (1) denoising of the scRNA-seq count matrix and (2) imputation based on the denoised data through a transcriptome-protein mapping (Figure 1a). The initial denoising, by SAVER-X [16], produces more accurate estimates of the RNA transcript relative abundances for each cell. Compared to the raw counts, the denoised relative expression values have significantly improved correlation with their corresponding protein measurement (Figure 1b, S3a, S4ab). Yet, for some genes, such as CD45RA, this correlation for denoised expression is still extremely low.

The production of a surface protein from its corresponding RNA transcript is a complicated process involving post-transcriptional modifications and transport [11], translation [12], post-translational modifications [13] and protein trafficking [14]. These processes depend on the state of the cell and the activities of other genes [9, 15]. To learn the mapping from a cell's transcriptome to the relative abundance for a given set of surface proteins, cTP-net employs a multiple branch deep neural network (MB-DNN, Figure S1). Deep neural networks have recently shown success in modeling complex biological systems [17, 18], and more importantly, allow good generalization across data sets [16, 19]. Generalization performance is an important aspect of cTP-net, as we

would like to perform imputation on tissues that do not exactly match the training data in cell type composition. Details of the cTP-net model and training procedure, as well as of alternative models and procedures that we have tried, are in Methods and Supplementary Note.

To examine imputation accuracy, we first consider the ideal case where imputation is conducted on cells of types that exactly match those in training data. For benchmarking, we used peripheral blood mononuclear cells (PBMCs) and cordical blood mononuclear cells (CBMCs) processed by CITE-seq and REAP-seq [2, 3], described in Table S1. We employed holdout method, where the cells in each data set were randomly partitioned into two sets: a training set with 90% of the cells and a holdout set with the remaining 10% of the cells for validation (Methods, Figure S2a). By law of large numbers, each cell type should be well represented in both the training and validation sets. Figure 1b and S3a show that, for all proteins examined in the CITE-seq PBMC data, cTP-net imputed abundances have higher correlation to the measured protein levels, as compared with the denoised and raw RNA counts of the corresponding genes. In most cases, the imputed values substantially improve upon the denoised values as proxies for protein abundance. We obtained similar results for the CITE-seq CBMC and REAP-seq PBMC data sets (Figure S4ab).

Next, we considered the generalization accuracy of cTP-net, testing whether it produces accurate imputations for cell types that are not present in the training set. For each of the high-level cell types in each data set in Table S2, all cells of the given type are held out during training, and cTP-net ,trained on the rest of the cells, was then used to impute protein abundances for the held out cells (Methods, Figure S2b). Across all benchmarking data sets and all cell types, these out-of-cell-type predictions are, as expected, inferior in accuracy to the traditional holdout validation predictions above, but still greatly improve upon the corresponding RNAs (Figure 2a, S4a). This indicates that cTP-net provides informative predictions on cell

types not present during training, vastly improving upon using the corresponding mRNA transcript abundance as proxy for the protein level.

To further examine the case where cell types in the training and test data are not perfectly aligned, we considered a scenario where the model is applied to perform imputation on a tissue that differs from the training data. We trained cTP-net on PBMCs and then applied it to perform imputation on CBMCs, and vice versa, using the data from Stoeckius et al. [3] (Methods). Cord blood is expected to be enriched for stem cells and cells undergoing differentiation, whereas peripheral blood contains well-differentiated cell types, and thus the two populations are composed of different but related cell types. Figure 2a and S3b shows the result on training on CBMC and then imputing on PBMC. As expected, imputing across tissue markedly improves the correlation to the measured protein level, as compared to the denoised RNA of the corresponding gene, but is worse than imputation produced by model trained on the same population. For practical use, we have trained a network using the two cell populations combined, which indeed achieves better accuracy than a network trained on each separately (Methods, Figure S3b, S4ac). The weights for this network are publicly available at https://github.com/zhouzilu/cTPnet.

We then tested whether cTP-net's predictions are sensitive to the laboratory protocol, and in particular, whether networks trained using CITE-seq data yields good predictions by REAP-seq's standard, and vice versa. Using a benchmarking design similar to above, we found that, in general, cTP-net maintains good generalization power across these two protocols (Figure 2a, S3b).

Seurat v3 anchor transfer [20] is a recent approach that uses cell alignment between data sets to impute features for single cell data. For comparison, we applied Seurat v3 anchor transfer to the holdout validation and out-of-cell-type benchmarking scenarios above (Methods). In the validation scenario, we found the performance of cTP-net and Seurat v3 to be comparable, with

cTP-net slightly better. cTP-net vastly improves upon Seurat in the out-of-cell-type scenario (Figure 2a, S5a). This is because cTP-net's neural network, trained across a diversity of cell types, learns a transcriptome-protein mapping that can more flexibly generalize to unseen cell types. As shown by the cross-population and out-of-cell-type benchmarking above, cTP-net does not require direct congruence of cell types across training and test sets.

So far, the correlations with measured protein abundance are computed across cells of all types. Since the proteins considered are highly cell type specific, such cross-type correlations are partly driven by the learning of cell type features, which are pronounced in the transcriptome data. Does cTP-net also capture the variation in protein abundance within the major cell types? As expected, within cell-type variation is harder to predict, but cTP-net's imputations nevertheless achieve high correlations with measured protein abundance for a subset of proteins and cell types (Figure S3c, S4d). Compared to Seurat v3, cTP-net's imputations align more accurately with measured protein levels when zoomed into cells of the same type (Figure 2b, S5b); see, for example, CD11c in CD14-CD16+ monocytes, CD2 in CD8 T cells, and CD19 in dendritic cells (Figure 2c). The learning of such within-type heterogeneity gives cTP-net the potential to attain higher resolution in the discovery and labeling of cell states.

What types of features are being used by cTP-net to form its imputation? To interpret the network, we conducted a permutation-based interpolation analysis, which calculates a prediction influence score for each protein-gene pair (Methods, Figure S6). Interpolation can be done using all cells, or cells of a specific type, the latter allowing us to probe relationships that may be specific to a given cell type. As expected, at the level of the general population that includes all cell types, the most influential genes for each protein are cell type specific genes (Table S3), since most of these surface proteins are cell type markers, and thus "cell type" is a key variable that underlies their heterogeneity. Within cell type interpolation, on the other hand, reveals genes related to RNA processing, RNA binding, protein localization and biosynthetic

process, in addition to immune-related genes that may be markers of cell sub-types (Table S4). This indicates that cTP-net combines different types of features, both cell type markers and genes involved in RNA to protein conversion and transport, to achieve multiscale imputation accuracy.

Having benchmarked cTP-net's generalization accuracy across immune cell types, tissues, and technologies, we then applied the network trained on the combined set of PBMCs and CBMCs from CITE-seq [3] to perform imputation for the Human Cell Atlas CBMC and bone marrow mononuclear cells (BMMC) data sets. Figure 2e shows the raw RNA count and predicted surface protein abundance for 10 markers across 7000 CBMCs in sample MantonCB2. (Similar plots for the other 7 CBMC and 8 BMMC samples are shown in Figure S8, S9). As expected based on the CITE-seq and REAP-seq studies, the imputed protein levels differ markedly from the RNA expression of its corresponding gene, displaying higher contrast across cell types and higher uniformity within cell type. The imputed protein levels can serve as intermediate features for the identification and labelling of cell states. For example, consider natural killer cells, in which proteins CD56 and CD16 serve as indicators for immunostimulatory effector functions, including an efficient cytotoxic capacity [21, 22]. We observe an opposing gradient of CD56 and CD16 levels within transcriptomically derived NK cell clusters that reveal $CD56^{bright}$ and $CD56^{dim}$ subsets, coherent with previous studies [3] (Figure 2f). This gradient in CD56 and CD16, where decrease in CD56 is accompanied by increase in CD16, is replicated across the 8 CBMC and 8 BMMC samples in HCA (Figure S8, S9). Consider also the case of CD57, which is a marker for terminally differentiated "senescent" cells in the T and NK cell types. The imputed level of CD57 is low, almost zero, in CBMCs, and rises in BMMCs. This is consistent with expectation since CD57+ NK cell and T cell populations grow after birth and with ageing [23-25] (Figure S8, S9). This shows that cTP-net, trained on a combination of CBMCs and PBMCs, can impute cell type and

cell stage specific protein signatures in new data without explicitly being given the tissue of origin.

Taken together, our results demonstrate that cTP-net can leverage existing CITE-seq and REAP-seq datasets to predict surface protein relative abundances for new scRNA-seq data sets, and the predictions generalize to cell types that are absent from, but related to those in the training data. Our benchmarking was done on diverse populations of PBMC and CBMC immune cells. With the accumulation of CITE-seq and REAP-seq data, cTP-net can be retrained to improve in accuracy and diversify in predictable protein targets. These results underscore the need for more diverse multi-omic cell atlases and demonstrate how such resources can be used to enhance future studies. The cTP-net package is available both in Python and R at https://github.com/zhouzilu/cTPnet.

Acknowledgements

Author Contributions

Z.Z. and N.Z. conceptualized the study and planned the case studies. Z.Z. designed the model, developed the algorithm, implemented the cTP-net software and led the data analysis. C.Y. helped in CITE-seq and REAP-seq data denoising and cell type labeling. J.W. helped with model design and Human Cell Atlas data analysis. Z.Z. and N.Z. wrote the paper with feedback from C.Y. and J.W.

Competing Financial Interests Statement

The authors declare no competing interests

References

1. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat Rev Genet* **20**, 257-272 (2019).
2. Peterson, V.M. et al. Multiplexed quantification of proteins and transcripts in single cells. *Nat Biotechnol* **35**, 936-939 (2017).
3. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single cells. *Nat Methods* **14**, 865-868 (2017).
4. Macaulay, I.C. et al. G&T-seq: parallel sequencing of single-cell genomes and transcriptomes. *Nat Methods* **12**, 519-522 (2015).
5. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional states. *Science* **361** (2018).
6. Regev, A. et al. The Human Cell Atlas. *Elife* **6** (2017).
7. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-cell RNA-seq. *Science* **352**, 189-196 (2016).
8. Villani, A.C. et al. Single-cell RNA-seq reveals new types of human blood dendritic cells, monocytes, and progenitors. *Science* **356** (2017).
9. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on mRNA Abundance. *Cell* **165**, 535-550 (2016).
10. Svensson, V. et al. Power analysis of single-cell RNA-sequencing experiments. *Nat Methods* **14**, 381-387 (2017).
11. Zhao, B.S., Roundtree, I.A. & He, C. Post-transcriptional gene regulation by mRNA modifications. *Nat Rev Mol Cell Biol* **18**, 31-42 (2017).
12. Jackson, R.J., Hellen, C.U. & Pestova, T.V. The mechanism of eukaryotic translation initiation and principles of its regulation. *Nat Rev Mol Cell Biol* **11**, 113-127 (2010).
13. Mowen, K.A. & David, M. Unconventional post-translational modifications in immunological signaling. *Nat Immunol* **15**, 512-520 (2014).
14. Schwartz, A.L. Cell biology of intracellular protein trafficking. *Annu Rev Immunol* **8**, 195-229 (1990).
15. Roux, P.P. & Topisirovic, I. Signaling Pathways Involved in the Regulation of mRNA Translation. *Mol Cell Biol* **38** (2018).
16. Wang, J. et al. Transfer learning in single-cell transcriptomics improves data denoising and pattern discovery. *bioRxiv*, 457879 (2018).
17. Webb, S. Deep learning for biology. *Nature* **554**, 555-557 (2018).
18. Tang, B., Pan, Z., Yin, K. & Khateeb, A. Recent Advances of Deep Learning in Bioinformatics and Computational Biology. *Front Genet* **10**, 214 (2019).
19. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nat Methods* **15**, 1053-1058 (2018).
20. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. *Cell* (2019).
21. Van Acker, H.H., Capsomidis, A., Smits, E.L. & Van Tendeloo, V.F. CD56 in the Immune System: More Than a Marker for Cytotoxicity? *Front Immunol* **8**, 892 (2017).
22. Tsukerman, P. et al. Expansion of CD16 positive and negative human NK cells in response to tumor stimulation. *Eur J Immunol* **44**, 1517-1525 (2014).
23. d'Angeac, A.D. et al. CD57+ T lymphocytes are derived from CD57- precursors by differentiation occurring in late immune responses. *Eur J Immunol* **24**, 1503-1511 (1994).
24. Musha, N. et al. Expansion of CD56+ NK T and gamma delta T cells from cord blood of human neonates. *Clin Exp Immunol* **113**, 220-228 (1998).
25. Dalle, J.H. et al. Characterization of cord blood natural killer cells: implications for transplantation and neonatal infections. *Pediatr Res* **57**, 649-655 (2005).
26. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444 (2015).
27. Kingma, D. & Ba, J. Adam: a method for stochastic optimization (2014). *arXiv preprint arXiv:1412.6980* **15** (2015).

28. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-15550 (2005).

Figure legends

Figure 1. cTP-net analysis pipeline and imputation of example proteins.

(a) Overview of cTP-net analysis pipeline, which learns a mapping from the denoised scRNA-seq data to the relative abundance of surface proteins, capturing multi-gene features that reflect the cellular environment and related processes. **(b)** For three example proteins, cross-cell scatter and correlation of CITE-seq measured abundances vs. (1) raw RNA count, (2) SAVER-X denoised RNA level, and (3) cTP-net predicted protein abundance.

Figure 2. Benchmark evaluation on CITE-seq PBMC data and imputation results on Human Cell Atlas CBMC data.

(a) Benchmark evaluation of cTP-net on CITE-seq PBMC data, with comparisons to Seurat v3, in validation, across cell type, across tissue and across technology scenarios. The table on the left shows the training scheme of each test, the heatmap shows correlations with actual measured protein abundances. **(b)** Within cell type correlations between imputed and measured protein abundance on the CITE-seq PBMC data, Seurat v3 versus cTP-net. Each point (color and shape pair) indicates a cell type and surface protein pair, where the x-axis is correlation between actual measured abundance and Seurat imputation and y-axis is the correlation between actual measured abundance and cTP-net imputation. **(c)** Scatter of imputed versus measured abundance for the three (surface protein, cell type) pairs marked by arrows in (b): CD11c in CD14-CD16+ monocytes, CD2 in CD8 T cells, and CD19 in dendritic cells. **(d)** t-SNE visualization of MantonCB2 CBMCs based on RNA expression, colored by cell type. B: B cells;

CD4 T: CD4 T cells; CD8 T: CD8 T cells; Mono: Monocyte; NK: Nature killer cells; Pre.: Precursors. **(e)** cTP-net imputed protein abundance and RNA read count of its corresponding gene for 12 surface proteins. **(f)** Enlarged plot of CD56 and CD16 imputed protein abundance and RNA read count in nature killer cells (NK). Reverse gradient is observed in cTP-net prediction but not in the read count for its corresponding RNA.

Online Methods

Data sets and pre-processing

Table S1 summarizes the five data sets analyzed in this study: CITE-PBMC, CITE-CBMC, REAP-PBMC, HCA-CBMC and HCA-BMMC. Among these, CITE-PBMC, CITE-CBMC and REAP-PBMC have paired scRNA-seq and surface protein counts, while HCA-CBMC and HCA-BMMC have only scRNA-seq counts. For all scRNA-seq data sets, low quality gene (< 10 counts across cells) and low quality cells (less than 200 genes detected) are removed, and the count matrix (C) for all remaining cells and genes is used as input for denoising. scRNA data denoising was performed with SAVER-X using default parameters. Denoised counts (Λ) were further transformed with Seurat default LogNormalize function,

$$X_{ij} = log\left(\frac{\Lambda_{ij} * 10{,}000}{m_j}\right)$$

where Λ_{ij} is the denoised molecule count of gene i in cell j, and m_j is the sum of all molecule counts of cell j. The normalized denoised count matrix X is the training input for the subsequent multiple branch neural network. For the surface protein counts, we adopted the relative abundance transformation from Stoeckius et al.[3]. For each cell c,

$$y_c = \left[\ln\left(\frac{p_{1c}}{g(p_c)}\right), \ln\left(\frac{p_{2c}}{g(p_c)}\right) \ldots \ln\left(\frac{p_{dc}}{g(p_c)}\right)\right]$$

where p_c is vector of antibody-derived tags (ADT) counts, and $g(p_c))$ is the geometric mean of p_c. The network trained using this transformed relative protein abundance as the response vector yields better prediction accuracy than the network trained using raw protein barcode counts.

cTP-net neural network structure and training parameters

Figure S1 shows the structure of cTP-net. Here, we have a normalized expression matrix X of N cells and D genes, and a normalized protein abundance matrix Y of the same N cells and d surface proteins. Let's denote cTP-net as a function F that maps from $\mathbb{R}^D$ to $\mathbb{R}^d$. Starting from the input layer, with dimension equals to number of genes D, the first internal layer has dimension 1000, followed by a second internal layer with dimension 128. These two layers are designed to learn and encode features that are shared across proteins, such as features that are informative for cell state and common processes such as cell cycle. The remaining layers are protein specific, with 64 nodes for each protein that feed into a one node output layer giving the imputed value. All layers except the last layer are fully connected (FC) with ReLU activation function [26], while the last layer is a fully connected layer with identity activation function for output. The objective function here is,

$$\underset{F}{\operatorname{argmin}} |Y - F(X)|_1^1$$

where the loss a L1 norm. The objective function was optimized stochastically with Adam [27]. Other variations of cTP-net, which we found to have inferior performance, are illustrated in more details in Supplementary Note.

Benchmarking procedure

Validation set testing procedure. Figure S2a shows the validation set testing procedure. Given limited amount of data, we keep only 10% of the cells as the testing set, and use the other 90% of the cells for training. The optimal model was selected based on the testing error.

Out-of-cell type prediction procedure. We perform the out-of-cell type prediction based on Figure S2b. This procedure mimics cross-validation, except that, instead of selecting the test set cells randomly, we partition the cells by their cell types. Iteratively, we designate all cells of a given cell type for testing and use the remaining cells for training. We then perform prediction on the hold-out cell type using the model trained on all other cell types. In the end, every cell has

been tested once and has the corresponding predictions. In the benchmark against the validation set testing procedure, we limit comparisons to the same cells that were in the validation set in the holdout scheme to account for variations between subsets.

Cell population and technology transfer learning procedure. To apply the models we trained in validation set testing procedure to different cell populations and technologies, the inputs have to be in the same feature space. Even though all data sets considered are from human cells, the list of genes differs between experiments and technologies. Genes that are in the training data but not in the testing data are filled with zeros. Because cTP-net utilizes overrepresented number of genes to predict the surface proteins level, having a small number of genes missing has little effect on the performance. After prediction, we selected only the shared proteins between two data sets for comparison.

cTP-net interpolation

To better interpret the relationships that the neural network is learning, we developed a permutation-based interpolation scheme that can calculate an influence score epi for each gene in the imputation of each protein (Figure S6). The idea is to assess how much changing the expression value of certain genes in the training data affects the training errors for a given model F. In each epoch, we interpolate all of the genes in a stochastic manner. As shown by Figure S6, the batch of genes denote by gs was randomly sampled. For genes within gs, cell labels were permuted. Here, the cell order within gs does not coordinate with protein abundance Y. We then calculate the prediction error (ϵ_z) on the interpolated cells (Z) and further compare with uninterpolated prediction error (ϵ_x). ϵ_{gs} is the influence score (relative importance) of gene set gs to this model F. We set batch size as 100 with 500 epochs. Furthermore, by picking different cells to interpolate, we could identify gene influence score in different cell types. For example, if matrix X belongs to a given cell type, the cell type specific genes are consistent

across cells of the given cell type, and thus, the permutation will not influence these genes. Genes that influence the surface protein abundance within the cell type, such as cell cycle genes and protein synthesis genes, tend to be rewarded with high influence scores in such a cell-type specific interpolation analysis.

For the top 100 highest influence scored genes from the following scenarios in CITE-PBMC: (1) CD45RA in CD14-CD16+ monocytes, (2) CD11c in CD14-CD16+ monocytes, (3) CD45RA in CD8 T cells, (4) CD45RA in CD4 T cells, (5) CD11c in CD14+CD16+ monocytes, (6) CD45RA in dendritic cells, and (7) CD11c in dendritic cells, we employed a Gene Ontology analysis [28] which identify top 10 pathways based on GO gene sets with FDR q-value < 0.05 as significant (Table S4).

Seurat anchor-transfer analysis

We compared cTP-net with an anchor-based transfer learning method developed in Seurat v3 [20]. For Seurat v3, RNA count data are normalized by LogNormalization method, while surface protein counts are normalized by centered log-ratio (CLR) method. In validation test setting, we used the same cells for training and testing as in cTP-net so as to be directly comparable to cTP-net. For out-of-cell type prediction, default parameters did not work for several cell types in anchor-transfer step, because, for those cell types, there are few anchors shared between the training and testing sets. To overcome this, we reduced the number of anchors iteratively until the function ran successfully.

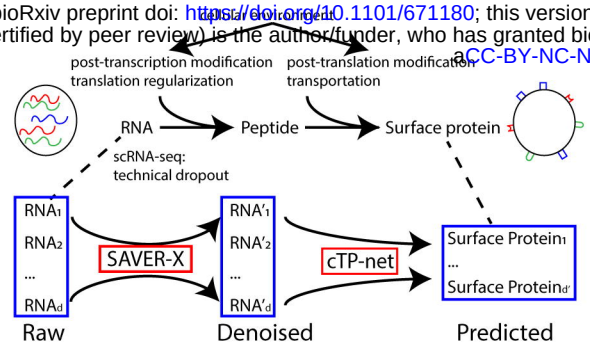
HCA data analysis

HCA RNA-seq transcriptome data analysis. HCA RNA-seq data sets are processed as discussed above, resulting in log-normalized denoised values. We applied default pipeline of Seurat and generated t-SNE plot for both data sets (Figure S7). Cells are clearly clustered by

individuals, indicating strong batch effects. As a result, the following analysis was performed on cells of each individual. Major cell types were determined by known gene markers.

Surface protein prediction by cTP-net. From the log-normalized denoised expression value, we predict the surface protein abundance with cTP-net model trained jointly on CITE-seq PBMC and CBMC data sets. We embedded 12 surface protein abundance across 16 individuals on t-SNE plot, showing consistent results with cell type information (Figure S8, S9).

a



b