bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

1 # Surface protein imputation from single cell transcriptomes
2 # by deep neural networks

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

4 1) Graduate Group in Genomics and Computational Biology, University of Pennsylvania,
5 Philadelphia, PA
6 2) Department of Statistics, University of Pennsylvania, Philadelphia, PA
7 3) School of Medicine, Tsinghua University, Beijing, China
8 * Correspondence:
9
10 Nancy R. Zhang
11 nzh@wharton.upenn.edu
12 (215) 898-8007
13 Department of Statistics
14 The Wharton School
15 University of Pennsylvania
16
17 **While single cell RNA sequencing (scRNA-seq) is invaluable for studying cell**
18 **populations, cell-surface proteins are often integral markers of cellular function and**
19 **serve as primary targets for therapeutic intervention. Here we propose a transfer learning**
20 **framework, single cell Transcriptome to Protein prediction with deep neural network**
21 **(cTP-net), to impute surface protein abundances from scRNA-seq data by learning from**
22 **existing single-cell multi-omic resources.**

23 Keywords: genomics, immunophenotypes, single cell sequencing, deep learning, prediction,
24 multi-omics

25 ## Introduction

26 Recent technological advances allow the simultaneous profiling, across many cells in parallel, of
27 multiple omics features in the same cell [1-5]. In particular, high throughput quantification of the
28 transcriptome and a selected panel of cell surface proteins in the same cell is now feasible

bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

29 through the REAP-seq and CITE-seq protocols [2, 3]. Cell surface proteins can serve as integral
30 markers of specific cellular functions and primary targets for therapeutic intervention.
31 Immunophenotyping by cell surface proteins has been an indispensable tool in hematopoiesis,
32 immunology and cancer research during the past 30 years. Yet, due to technological barriers
33 and cost considerations, most single cell studies, including Human Cell Atlas project [6], quantify
34 the transcriptome only and do not have cell-matched measurements of relevant surface proteins
35 [7, 8]. Sometimes, which cell types and corresponding surface proteins are essential become
36 apparent only after exploration by scRNA-seq. This motivates our inquiry of whether protein
37 abundances in individual cells can be accurately imputed by the cell's transcriptome.

38 We propose cTP-net (single cell Transcriptome to Protein prediction with deep neural network),
39 a transfer learning approach based on deep neural networks that imputes surface protein
40 abundances for scRNA-seq data. Through comprehensive benchmark evaluations and
41 applications to Human Cell Atlas and acute myeloid leukemia data sets, we show that cTP-net
42 outperform existing methods and can transfer information from training data to accurately
43 impute 24 immunophenotype markers, which achieve a more detailed characterization of
44 cellular state and cellular phenotypes than transcriptome measurements alone. cTP-net relies,
45 for model training, on accumulating public data of cells with paired transcriptome and surface
46 protein measurements.

47

48 Results

49 Method overview

50 An overview of cTP-net is shown in Figure 1a. Studies based on both CITE-seq and REAP-seq
51 have shown that the relative abundance of most surface proteins, at the level of individual cells,

bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

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

63 The production of a surface protein from its corresponding RNA transcript is a complicated
64 process involving post-transcriptional modifications and transport [11], translation [12], post-
65 translational modifications [13] and protein trafficking [14]. These processes depend on the state of
66 the cell and the activities of other genes [9, 15]. To learn the mapping from a cell's transcriptome to
67 the relative abundance of a given set of surface proteins, cTP-net employs a multiple branch
68 deep neural network (MB-DNN, Figure S1). Deep neural networks have recently shown success
69 in modeling complex biological systems [17, 18], and more importantly, allow good generalization
70 across data sets[16, 19]. Generalization performance is an important aspect of cTP-net, as we
71 would like to perform imputation on tissues that do not exactly match the training data in cell
72 type composition. Details of the cTP-net model and training procedure, as well as of alternative
73 models and procedures that we have tried, are in Methods and Supplementary Note.

74

75 Imputation accuracy assessment and transfer learning

bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

76 ## Evaluation via random holdout

77 To examine imputation accuracy, we first consider the ideal case where imputation is conducted
78 on cells of types that exactly match those in training data. For benchmarking, we used
79 peripheral blood mononuclear cells (PBMCs) and cord blood mononuclear cells (CBMCs)
80 processed by CITE-seq and REAP-seq [2, 3], described in Table S1. We employed holdout
81 method, where the cells in each data set were randomly partitioned into two sets: a training set
82 with 90% of the cells and a holdout set with the remaining 10% of the cells for validation
83 (Methods, Figure S2a). Each cell type is well represented in both the training and validation
84 sets. Figure 1b and S3a show that, for all proteins examined in the CITE-seq PBMC data, cTP-
85 net imputed abundances have much higher correlation to the measured protein levels, as
86 compared with the denoised and raw RNA counts of the corresponding genes. We obtained
87 similar results for the CITE-seq CBMC and REAP-seq PBMC data sets (Figure S4ab).

88

89 ## Generalization to unseen cell types

90 Next, we considered the generalization accuracy of cTP-net, testing whether it produces
91 accurate imputations for cell types that are not present in the training set. For each of the high-
92 level cell types in each data set in Table S2, all cells of the given type are held out during
93 training, and cTP-net, trained on the rest of the cells, was then used to impute protein
94 abundances for the held out cells (Methods, Figure S2b). We did this for each cell type and
95 generated an “out-of-cell-type” prediction for every cells.

96 Across all benchmarking data sets and all cell types, these out-of-cell-type predictions still
97 improve significantly upon the corresponding RNA counts while slightly inferior in accuracy to
98 the traditional holdout validation predictions above (Figure 2a, S4a). This indicates that cTP-net

bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

99 provides informative predictions on cell types not present during training, vastly improving upon
100 using the corresponding mRNA transcript abundance as proxy for the protein level.

101

102 Generalization across tissue and lab protocol

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

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

121

122 Comparison to Seurat v3

bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

123 Seurat v3 anchor transfer [20] is a recent approach that uses cell alignment between data sets to
124 impute features for single cell data. For comparison, we applied Seurat v3 anchor transfer to the
125 holdout validation and out-of-cell-type benchmarking scenarios above (Methods). In the
126 validation scenario, we found the performance of cTP-net and Seurat v3 to be comparable, with
127 cTP-net slightly better, as both methods can estimate protein abundance by utilizing marker
128 genes to identify the cell types. cTP-net, however, vastly improves upon Seurat in the out-of-
129 cell-type scenario (Figure 2a, S5a). This is because cTP-net's neural network, trained across a
130 diversity of cell types, learns a direct transcriptome-protein mapping that can more flexibly
131 generalize to unseen cell types, while Seurat v3 depends on a nearest neighbor method that
132 can only sample from the training dataset. As shown by the cross-population and out-of-cell-
133 type benchmarking above, cTP-net does not require direct congruence of cell types across
134 training and test sets.

135 In addition to predictions on unseen cell type, cTP-net also improves upon the existing state-of-
136 the-art in capturing within cell-type variation in protein abundance. As expected, within cell-type
137 variation is harder to predict, but cTP-net's imputations nevertheless achieve high correlations
138 with measured protein abundance for a subset of proteins and cell types (Figure S3c, S4d).
139 Compared to Seurat v3, cTP-net's imputations align more accurately with measured protein
140 levels when zoomed into cells of the same type (Figure 2b, S5b); see Figure 2c, for example,
141 CD11c in CD14-CD16+ monocytes, CD2 in CD8 T cells, and CD16 in dendritic cells. All of these
142 surface proteins have important biological function in the corresponding cell types, as CD11c
143 helps trigger respiratory burst in monocyte [21], CD2 co-stimulates molecule on T cells [22] and
144 CD16 differentiate DC subpopulation [23]. The learning of such within-type heterogeneity gives
145 cTP-net the potential to attain higher resolution in the discovery and labeling of cell states.

146

147 Network interpretation and feature importance

bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

148 What types of features are being used by cTP-net to form its imputation? To interpret the
149 network, we conducted a permutation-based interpolation analysis, which calculates a
150 permutation feature importance for each protein-gene pair (Methods, Figure S6a). Interpolation
151 can be done using all cells, or cells of a specific type, the latter allowing us to probe
152 relationships that may be specific to a given cell type. Applying this analysis to cTP-net trained
153 on PBMC, we found that, at the level of the general population that includes all cell types, the
154 most important genes for the prediction of each protein are those that exhibit the highest cell-
155 type specificity in expression (Table S3). This is because most of these surface proteins are
156 cell type markers, and thus when cells of all types are pooled together, "cell type" is the key
157 latent variable that underlies their heterogeneity. Within cell type interpolation, on the other
158 hand, reveals genes related to RNA processing, RNA binding, protein localization and
159 biosynthetic processes, in addition to immune-related genes that differentiate the immune cell
160 sub-types (Table S4). This analysis shows that cTP-net combines different types of features,
161 both cell type markers and genes involved in RNA to protein conversion and transport, to
162 achieve multiscale imputation accuracy.

163 In addition, we analyzed the bottleneck layer with 128 nodes before the network branched out to
164 the protein-specific layers. We performed dimension reduction (UMAP) directly on the
165 bottleneck layer intermediate output of 7000 PBMCs from CITE-seq. Figure S6b shows that the
166 cells are cleanly separated into different clusters, representing cell types as well as gradients in
167 surface protein abundance. This confirms that the bottleneck layer captures the essential
168 information on cell stages and transitions, and that each subsequent individual branch then
169 predicts its corresponding protein's abundance.

170

171 Application to Human Cell Atlas

bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

172 Having benchmarked cTP-net's generalization accuracy across immune cell types, tissues, and
173 technologies, we then applied the network trained on the combined CITE-seq dataset of
174 PBMCs,CBMCs and bone marrow mononuclear cells (BMMCs) [3, 24] to perform imputation for the
175 Human Cell Atlas CBMC and BMMC data sets (Table S1). Figure 3 shows the raw RNA count
176 and predicted surface protein abundance for 24 markers across 6023 BMMCs from sample
177 MantonBM1 and 4176 CBMCs from sample MantonCB1. (Similar plots for the other 7 BMMC
178 and 7 CBMC samples are shown in Figure S8, S9). Similar to what was observed for actual
179 measured protein abundances in the CITE-seq and REAP-seq studies, the imputed protein
180 levels differ markedly from the RNA expression of its corresponding gene, displaying higher
181 contrast across cell types and higher uniformity within cell type. Thus, the imputed protein levels
182 can serve as interpretable intermediate features for the identification and labelling of cell states.
183 For example, imputed CD4 and CD8 levels separate CD4+ T cells from CD8+ T cells with high
184 confidence. Further separation of naïve T cells to memory T cells can be achieved through
185 imputed CD45RA/CD45RO abundance, as CD45RA is a naïve antigen and CD45RO is a
186 memory antigen. Consistent with flow cytometry data, the large majority of CB T cells are naïve,
187 whereas the BM T cell population is more diverse [25]. Also, for BM B cells that have high imputed
188 CD19 levels, cTP-net allows us to confidently distinguish the Pre.B (CD38+, CD127+), immature
189 B (CD38+, CD79b+), memory B (CD27+) and naïve B cells (CD27-), whose immunophenotypes
190 have been well characterized [26].

191 In addition, consider natural killer cells, in which the proteins CD56 and CD16 serve as
192 indicators for immunostimulatory effector functions, including an efficient cytotoxic capacity [27, 28].
193 We observe an opposing gradient of imputed CD56 and CD16 levels within transcriptomically
194 derived natural killer (NK) cell clusters that reveal $CD56^{bright}$ and $CD56^{dim}$ subsets, coherent with
195 previous studies[3] (Figure 2f, Figure S10, F-test: p-value = 1.667e-15). This pattern is not found
196 in RNA abundances due to low expression (F-test: p-value= 0.9377). Between $CD56^{brignt}$ and

bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

197 CD56^{dim} subsets, 7 out of 10 of previously studied differentially expressed genes are significant
198 in the single cell analysis (Fisher test: p-value = 1.07e-04) [3, 29, 30]. This gradient in CD56 and
199 CD16, where decrease in CD56 is accompanied by increase in CD16, is replicated across the 8
200 CBMC and 8 BMMC samples in HCA (Figure S8, S9, S10).

201 Consider also the case of CD57, which is a marker for terminally differentiated "senescent" cells
202 in the T and NK cell types. The imputed level of CD57 is lower in CBMCs (fetus's blood), and
203 rises in BMMCs (95% quantile: bootstrap p-value<1e-6). This is consistent with expectation
204 since CD57+ NK cell and T cell populations grow after birth and with ageing [31-33] (Figure S8,
205 S9).

206 These results demonstrate how cTP-net, trained on a combination of PBMCs, CBMCs and
207 BMMCs, can impute cell type, cell stage, and tissue-specific protein signatures in new data
208 without explicitly being given the tissue of origin.

209

210 Application to Acute Myeloid Leukemia

211 We further apply cTP-net to an acute myeloid leukemia (AML) data set from Galen et al. [34]. AML
212 is a heterogeneous disease where the diversity of malignant cell types partially recapitulates the
213 stages of myeloid development. Mapping the malignant cells in AML to the differentiation stage
214 of their cell of origin strongly impacts tumor prognosis and treatment, as malignant cells that
215 originate from earlier stage progenitors have higher risk of relapse [35, 36]. In the original paper,
216 the authors sequenced 7698 cells from 5 healthy donors to build a reference map of cell types
217 during myeloid development, and then mapped 30712 cells from 16 AML patients across
218 multiple time points to this reference to identify the differentiation stage of the malignant cells.
219 Here, by imputing 24 immunophenotype markers with cTP-net, we can directly characterize the
220 differentiation stage of cell-of-origin for the malignant cells.

bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

221 Figure 4a is a UMAP plot based on imputed surface protein abundance of 5 normal BMs and 12
222 Day 0 samples from AML patients. The majority of the malignant cells as identified in the
223 original paper reside on the right half of the plot, which recapitulate the myeloid differentiation
224 trajectory as revealed by the imputed values of canonical protein markers (Figure 4b): From
225 CD34+ progenitors to CD38+CD123+ cells in transition to CD11c+ and CD14+ mature
226 monocytes [37]. All of the malignant cells have imputed protein values that place them along this
227 monocyte lineage. Using the transcriptome for visualization, on the other hand, reveals large
228 batch effects across samples, due to both technical batch and biological differences (Figure
229 S11). Thus, unlike the imputed protein data, the transcriptomic data cannot be directly
230 combined without alignment.

231 Based on the trajectory revealed by the imputed protein levels, we can determine the
232 differentiation cell stage(s) for the malignant cells of each tumor, according to which the 12 AML
233 patients can be divided into three categories: (1) AMLs of single differentiation stage (AML420B,
234 AML556, AML707B and AML916; Figure 4c), (2) AMLs of two differentiation stages (AML210A,
235 AML328, AML419A and AML475; Figure 4e) and (3) AMLs of many differentiation stages
236 (AML1012, AML329, AML870 and AML921A; Figure 4f). This stage assignment is consistent
237 with the original study [34]. For example, AML419A harbors two malignant cell types at opposite
238 ends of the monocyte differentiation axis, distinguished by imputed CD34 and CD11c levels as
239 CD34+CD11c- indicates progenitor-like and CD34-CD11c+ indicates differentiated monocyte-
240 like cells (Figure 4d, 4e). AML707B, which carries a RUNX1/RUNX1T1 fusion, consists of cells
241 of a specific cell stage that is distinct from the normal myeloid trajectory (Figure 4c). Such
242 unique cell cluster was due to hyper CD38 level in surface protein prediction (Figure 4d). Such
243 hyper-CD38 levels have been reported in AMLs with RUNX1/RUNX1T1 fusion[38-40] and recent
244 studies have also shown that CD38 can be a potential target for adult AML[41, 42].

bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

245 In this example, the imputed protein levels served as useful features for trajectory visualization.
246 This analysis also indicates that even though cTP-net is currently trained only on normal
247 immune cells, it can reveal disease-specific signatures in malignant cells and the imputed
248 protein levels are useful for characterizing tumor phenotypes.

249

250 Discussion

251 Taken together, our results demonstrate that cTP-net can leverage existing CITE-seq and
252 REAP-seq datasets to predict surface protein relative abundances for new scRNA-seq data
253 sets, and that the predictions generalize to cell types that are absent from, but related to those
254 in the training data. cTP-net was benchmarked on PBMC and CBMC immune cells, showing
255 good generalization across tissues and technical protocols. On Human Atlas Data, we show
256 that the imputed surface protein levels allow easy assignment of cells to known cell types, as
257 well as the revealing of intra-cell type gradients. We then demonstrate that, even though cTP-
258 net used only immune cells from healthy individuals for training, it is able to impute
259 immunophenotypes for malignant cells from acute myeloid leukemia, and that these
260 immunophenotypes allow placement of the cells along the myeloid differentiation trajectory.
261 Furthermore, we show that cTP-net is able to impute protein signatures in the malignant cells
262 that are disease specific and that are not easily detectable from the transcriptomic counts.

263 SAVER-X serves an important role in the training procedure of cTP-net. As shown in Table S5,
264 without SAVER-X denoising, the cTP-net prediction performance retracts by 0.02 in correlation,
265 more significant than any other parameter tweaks. This discrepancy in performance is due to:
266 (1) SAVER-X makes use of the noise model to obtain estimates of the true RNA counts. This
267 helps cTP-net learn the underlining relationship between true RNA counts and protein level,
268 rather than the noisy raw counts and protein levels, which varies more across data sets and

bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

269 thus does not generalize well. (2) By denoising the scRNA-seq, the input for learning the RNA-
270 protein relationship is less sparse. Manifold learning on a more continuous input space usually
271 works better[43, 44]. (3) Comparing to other autoencoder based denoising method, SAVER-X
272 performs Bayesian shrinkage on top of autoencoder framework to prevent over-imputation
273 (over-smoothing) [16, 45].

274 Despite these promising results, cTP-net has limitations. (1) cTP-net can only apply to count
275 based expression input (UMI-based). CITE-seq data with TPM and RPKM expression metric is
276 not available for testing. Thus, the prediction accuracy is unknown. (2) The generalization ability
277 of cTP-net to unrelated cell types has limitations. Even though the final cTP-net model, trained
278 on immune cells, has good results on immune cells from diverse settings, we have not tried to
279 perform imputation of these immune-related markers on cells that are not of the hematopoietic
280 lineage.

281 With the accumulation of publicly available CITE-seq and REAP-seq data across diverse
282 proteins, cell types and conditions, cTP-net can be retrained to accommodate more protein
283 targets and improve in generalization accuracy. The possibility of such cross-omic transfer
284 learning underscores the need for more diverse multi-omic cell atlases, and demonstrate how
285 such resources can be used to enhance future studies. The cTP-net package is available both
286 in Python and R at https://github.com/zhouzilu/cTPnet.

287

288 ## Acknowledgements

289 We would like to thank the National Institutes of Health for the award 5R01-HG006137 (for Z.Z.,
290 and N.Z.), award 1U2CCA233285-01 (for N.Z.), the National Science Foundation for the

bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

291 award DMS-1562665 (to J.W., N.Z.), and the Wharton Dean's Fund for Post-doctoral
292 Research (to J.W.).

293 ## Author Contributions

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

299 ## Competing Financial Interests Statement

300 The authors declare no competing interests

301

302 ## References

303 1. Stuart, T. & Satija, R. Integrative single-cell analysis. *Nat Rev Genet* **20**, 257-272 (2019).
304 2. Peterson, V.M. et al. Multiplexed quantification of proteins and transcripts in single cells.
305 *Nat Biotechnol* **35**, 936-939 (2017).
306 3. Stoeckius, M. et al. Simultaneous epitope and transcriptome measurement in single
307 cells. *Nat Methods* **14**, 865-868 (2017).
308 4. Macaulay, I.C. et al. G&T-seq: parallel sequencing of single-cell genomes and
309 transcriptomes. *Nat Methods* **12**, 519-522 (2015).
310 5. Wang, X. et al. Three-dimensional intact-tissue sequencing of single-cell transcriptional
311 states. *Science* **361** (2018).
312 6. Regev, A. et al. The Human Cell Atlas. *Elife* **6** (2017).
313 7. Tirosh, I. et al. Dissecting the multicellular ecosystem of metastatic melanoma by single-
314 cell RNA-seq. *Science* **352**, 189-196 (2016).
315 8. Villani, A.C. et al. Single-cell RNA-seq reveals new types of human blood dendritic cells,
316 monocytes, and progenitors. *Science* **356** (2017).
317 9. Liu, Y., Beyer, A. & Aebersold, R. On the Dependency of Cellular Protein Levels on
318 mRNA Abundance. *Cell* **165**, 535-550 (2016).
319 10. Svensson, V. et al. Power analysis of single-cell RNA-sequencing experiments. *Nat*
320 *Methods* **14**, 381-387 (2017).
321 11. Zhao, B.S., Roundtree, I.A. & He, C. Post-transcriptional gene regulation by mRNA
322 modifications. *Nat Rev Mol Cell Biol* **18**, 31-42 (2017).

bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

323 12. Jackson, R.J., Hellen, C.U. & Pestova, T.V. The mechanism of eukaryotic translation
324 initiation and principles of its regulation. *Nat Rev Mol Cell Biol* **11**, 113-127 (2010).
325 13. Mowen, K.A. & David, M. Unconventional post-translational modifications in
326 immunological signaling. *Nat Immunol* **15**, 512-520 (2014).
327 14. Schwartz, A.L. Cell biology of intracellular protein trafficking. *Annu Rev Immunol* **8**, 195-
328 229 (1990).
329 15. Roux, P.P. & Topisirovic, I. Signaling Pathways Involved in the Regulation of mRNA
330 Translation. *Mol Cell Biol* **38** (2018).
331 16. Wang, J. et al. Data denoising with transfer learning in single-cell transcriptomics. *Nat*
332 *Methods* **16**, 875-878 (2019).
333 17. Webb, S. Deep learning for biology. *Nature* **554**, 555-557 (2018).
334 18. Tang, B., Pan, Z., Yin, K. & Khateeb, A. Recent Advances of Deep Learning in
335 Bioinformatics and Computational Biology. *Front Genet* **10**, 214 (2019).
336 19. Lopez, R., Regier, J., Cole, M.B., Jordan, M.I. & Yosef, N. Deep generative modeling for
337 single-cell transcriptomics. *Nat Methods* **15**, 1053-1058 (2018).
338 20. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. *Cell* (2019).
339 21. Martins, P.S. et al. Expression of cell surface receptors and oxidative metabolism
340 modulation in the clinical continuum of sepsis. *Crit Care* **12**, R25 (2008).
341 22. Chen, L. & Flies, D.B. Molecular mechanisms of T cell co-stimulation and co-inhibition.
342 *Nat Rev Immunol* **13**, 227-242 (2013).
343 23. Fromm, P. et al. CD16+Dendritic Cells Are a Unique Myeloid Antigen Presenting Cell
344 Population. *Blood* **128** (2016).
345 24. Stuart, T. et al. Comprehensive Integration of Single-Cell Data. *Cell* **177**, 1888-+ (2019).
346 25. D'Arena, G. et al. Flow cytometric characterization of human umbilical cord blood
347 lymphocytes: immunophenotypic features. *Haematologica* **83**, 197-203 (1998).
348 26. Clavarino, G. et al. Novel Strategy for Phenotypic Characterization of Human B
349 Lymphocytes from Precursors to Effector Cells by Flow Cytometry. *Plos One* **11** (2016).
350 27. Van Acker, H.H., Capsomidis, A., Smits, E.L. & Van Tendeloo, V.F. CD56 in the Immune
351 System: More Than a Marker for Cytotoxicity? *Front Immunol* **8**, 892 (2017).
352 28. Tsukerman, P. et al. Expansion of CD16 positive and negative human NK cells in
353 response to tumor stimulation. *Eur J Immunol* **44**, 1517-1525 (2014).
354 29. Poli, A. et al. CD56(bright) natural killer (NK) cells: an important NK cell subset.
355 *Immunology* **126**, 458-465 (2009).
356 30. Wendt, K. et al. Gene and protein characteristics reflect functional diversity of CD56(dim)
357 and CD56(bright) NK cells. *J Leukocyte Biol* **80**, 1529-1541 (2006).
358 31. d'Angeac, A.D. et al. CD57+ T lymphocytes are derived from CD57- precursors by
359 differentiation occurring in late immune responses. *Eur J Immunol* **24**, 1503-1511 (1994).
360 32. Musha, N. et al. Expansion of CD56+ NK T and gamma delta T cells from cord blood of
361 human neonates. *Clin Exp Immunol* **113**, 220-228 (1998).
362 33. Dalle, J.H. et al. Characterization of cord blood natural killer cells: implications for
363 transplantation and neonatal infections. *Pediatr Res* **57**, 649-655 (2005).
364 34. van Galen, P. et al. Single-Cell RNA-Seq Reveals AML Hierarchies Relevant to Disease
365 Progression and Immunity. *Cell* **176**, 1265-+ (2019).
366 35. Pollyea, D.A. & Jordan, C.T. Therapeutic targeting of acute myeloid leukemia stem cells.
367 *Blood* **129**, 1627-1635 (2017).
368 36. McKenzie, M.D. et al. Interconversion between Tumorigenic and Differentiated States in
369 Acute Myeloid Leukemia. *Cell Stem Cell* **25**, 258-+ (2019).
370 37. Geissmann, F. et al. Development of Monocytes, Macrophages, and Dendritic Cells.
371 *Science* **327**, 656-661 (2010).

bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

372 38. Jang, J.H. et al. Acute myeloid leukemia with del(X)(p21) and cryptic RUNX1/RUNX1T1
373 from ins(8;21)(q22;q22q22) revealed by atypical FISH signals. *Ann Clin Lab Sci* **40**, 80-
374 84 (2010).
375 39. Moroi, K. & Sato, T. Comparison between procaine and isocarboxazid metabolism in
376 vitro by a liver microsomal amidase-esterase. *Biochem Pharmacol* **24**, 1517-1521
377 (1975).
378 40. Shang, L. et al. The immunophenotypic characteristics and flow cytometric scoring
379 system of acute myeloid leukemia with t(8;21) (q22;q22); RUNX1-RUNX1T1. *Int J Lab*
380 *Hematol* **41**, 23-31 (2019).
381 41. Naik, J. et al. CD38 as a therapeutic target for adult acute myeloid leukemia and T-cell
382 acute lymphoblastic leukemia. *Haematologica* **104**, E100-E103 (2019).
383 42. Eveillard, M. et al. CD38 Expression in B-Lineage Acute Lymphoblastic Leukemia, a
384 Possible Target for Immunotherapy. *Blood* **128** (2016).
385 43. An, G.Z. The effects of adding noise during backpropagation training on a generalization
386 performance. *Neural Comput* **8**, 643-674 (1996).
387 44. Reed, R. & MarksII, R.J. Neural smithing: supervised learning in feedforward artificial
388 neural networks. (Mit Press, 1999).
389 45. Andrews, T.S. & Hemberg, M. False signals induced by single-cell imputation.
390 *F1000Res* **7**, 1740 (2018).
391 46. LeCun, Y., Bengio, Y. & Hinton, G. Deep learning. *Nature* **521**, 436-444 (2015).
392 47. Kingma, D. & Ba, J. Adam: a method for stochastic optimization (2014). *arXiv preprint*
393 *arXiv:1412.6980* **15** (2015).
394 48. Subramanian, A. et al. Gene set enrichment analysis: a knowledge-based approach for
395 interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* **102**, 15545-
396 15550 (2005).

397

398

bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

399 # Figure legends

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

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

406 Figure 2. Benchmark evaluation on CITE-seq PBMC data

407 **(a)** Benchmark evaluation of cTP-net on CITE-seq PBMC data, with comparisons to Seurat v3,
408 in validation, across cell type, across tissue and across technology scenarios. The table on the
409 left shows the training scheme of each test, the heatmap shows correlations with actual
410 measured protein abundances. **(b)** Within cell type correlations between imputed and measured
411 protein abundance on the CITE-seq PBMC data, Seurat v3 versus cTP-net. Each point (color
412 and shape pair) indicates a cell type and surface protein pair, where the x-axis is correlation
413 between actual measured abundance and Seurat imputation and y-axis is the correlation
414 between actual measured abundance and cTP-net imputation. **(c)** Scatter of imputed versus
415 measured abundance for the three (surface protein, cell type) pairs marked by arrows in (b):
416 CD11c in CD14-CD16+ monocytes, CD2 in CD8 T cells, and CD19 in dendritic cells.

417 Figure 3. Imputation results analysis on Human Cell Atlas data sets.

418 **(a)** Left panel: UMAP visualization of MantonBM1 BMMCs T cell subpopulation based on RNA
419 expression, colored by cell type. CD4 T: mature CD4+ T cells; mature CD8 T: CD8+ T cells;
420 naïve CD4 T: naïve CD4+ T cells; naïve CD8 T: naïve CD8+ T cells; CD8 senescent T: CD8+
421 senescent T cells. Right panel: Related imputed protein abundance and RNA expression of its

bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

422 corresponding gene. **(b)** UMAP visualization of MantonBM1 BMMCs based on RNA expression,
423 colored by cell type. B: B cells; CD4 T: CD4+ T cells; CD8 T: CD8+ T cells; cMono: classical
424 monocyte; ncMono: non-classical monocyte; NK: natural killer cells; Pre.: precursors; Plasma:
425 plasma cells. **(c)** Left panel: UMAP visualization of MantonBM1 BMMCs B cell subpopulation
426 based on RNA expression, colored by cell type. Pre.B: B cell precursors; immature B: immature
427 B cells; memory B: memory B cells; naïve B: naïve B cells. Right panel: Related imputed protein
428 abundance and RNA expression of its corresponding gene. **(d)** UMAP visualization of
429 MantonCB2 CBMCs based on RNA expression, colored by cell type. **(e)** cTP-net imputed
430 protein abundance and RNA read count of its corresponding gene for 24 surface proteins. **(f)**
431 UMAP visualization of MantonCB2 CBMCs NK cell subpopulation colored by CD56 and CD16
432 imputed protein abundance and RNA read count. Reverse gradient is observed in cTP-net
433 prediction but not in the read count for its corresponding RNA. **(f)** Contour plot of cells based on
434 imputed CD56 and CD16 abundance in NK cell populations. Strong negative correlation with
435 two subpopulation observed.

436 Figure 4. Imputation results analysis on Acute Myeloid Leukemia data sets.

437 **(a)** UMAP visualization of normal cells and malignant cells from 12 AML samples at Day0 based
438 on imputed protein abundance (red: malignant cells; grey: normal cells). **(b)** UMAP visualization
439 of the myeloid trajectory. cTP-net imputed protein abundance of markers that perfectly
440 recapitulate the myeloid development. **(c, e, f)** UMAP visualization of the myeloid trajectory with
441 corresponding malignant cells from AML sample highlighted. **(d)** Scatter plot of normal and AML
442 malignant cells based on imputed protein expression.

443

bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

444 # Online Methods

445 ## Data sets and pre-processing

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

454 $$X_{ij} = log\left(\frac{\Lambda_{ij} * 10{,}000}{m_j}\right)$$

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

459 $$y_c = \left[\ln\left(\frac{p_{1c}}{g(p_c)}\right), \ln\left(\frac{p_{2c}}{g(p_c)}\right) \ldots \ln(\frac{p_{dc}}{g(p_c)})\right]$$

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

464

bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

465 **cTP-net neural network structure and training parameters**

466 Figure S1 shows the structure of cTP-net. Here, we have a normalized expression matrix X of N
467 cells and D genes, and a normalized protein abundance matrix Y of the same N cells and d
468 surface proteins. Let's denote cTP-net as a function F that maps from $\mathbb{R}^D$ to $\mathbb{R}^d$. Starting from
469 the input layer, with dimension equals to number of genes D, the first internal layer has
470 dimension 1000, followed by a second internal layer with dimension 128. These two layers are
471 designed to learn and encode features that are shared across proteins, such as features that
472 are informative for cell type, cell state and common processes such as cell cycle. The remaining
473 layers are protein specific, with 64 nodes for each protein that feed into a one node output layer
474 giving the imputed value. All layers except the last layer are fully connected (FC) with rectified
475 linear unit (ReLU) activation function [46], while the last layer is a fully connected layer with
476 identity activation function for output. The objective function here is,

477 $$\underset{F}{\operatorname{argmin}} |Y - F(X)|_1$$

478 where the loss is L1 norm. The objective function was optimized stochastically with Adam [47] with
479 learning rate set to 10e-5 for 139 epochs (cross-validation). Other variations of cTP-net, which
480 we found to have inferior performance, are illustrated in more details in Supplementary Note.

481

482 **Benchmarking procedure**

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

486 *Out-of-cell type prediction procedure.* We perform the out-of-cell type prediction based on
487 Figure S2b. This procedure mimics cross-validation, except that, instead of selecting the test set

bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

488 cells randomly, we partition the cells by their cell types. Iteratively, we designate all cells of a
489 given cell type for testing and use the remaining cells for training. We then perform prediction on
490 the hold-out cell type using the model trained on all other cell types. In the end, every cell has
491 been tested once and has the corresponding predictions. In the benchmark against the
492 validation set testing procedure, we limit comparisons to the same cells that were in the
493 validation set in the holdout scheme to account for variations between subsets.

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

502

503 **cTP-net interpolation**

504 To better interpret the relationships that the neural network is learning, we developed a
505 permutation-based interpolation scheme that can calculate an influence score epi for each gene
506 in the imputation of each protein (Figure S6). The idea is to assess how much changing the
507 expression value of certain genes in the training data affects the training errors for a given
508 model F. In each epoch, we interpolate all of the genes in a stochastic manner. Let's denote X
509 as the expression matrix (N by G matrix, where N is the number of cells and G is number of
510 genes), Y as protein abundance matrix and L as the loss function. The algorithm goes as follow
511 (Figure S6):

bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

512 (1) Estimate the original model error $\epsilon^{orig} = L(Y, F(X))$.
513 (2) Sampling batch of genes denote by gs . Generate expression matrix X^{perm} by permuting
514 genes in gs in the data X. This breaks the association between gs and protein
515 abundance Y, i.e. the cell order within gs does not coordinate with protein abundance Y.
516 (3) Estimate error $\epsilon^{perm} = L(Y, F(X^{perm}))$ based on the predictions of the permuted data.
517 (4) Calculate permutation feature importance $\Delta_{gs} = |\epsilon^{orig} - \epsilon^{perm}|$ of gene set gs to this
518 model F.

519 We set batch size as 100 with 500 epochs. Furthermore, by picking different cells to interpolate,
520 we could identify gene influence score in different cell types. For example, if matrix X belongs to
521 a given cell type, the cell type specific genes are consistent across cells of the given cell type,
522 and thus, the permutation will not influence these genes. Genes that influence the surface
523 protein abundance within the cell type, such as cell cycle genes and protein synthesis genes,
524 tend to be rewarded with high influence scores in such a cell-type specific interpolation analysis.

525 For the top 100 highest influence scored genes from the following scenarios in CITE-PBMC: (1)
526 CD45RA in CD14-CD16+ monocytes, (2) CD11c in CD14-CD16+ monocytes, (3) CD45RA in
527 CD8 T cells, (4) CD45RA in CD4 T cells, (5) CD11c in CD14+CD16+ monocytes, (6) CD45RA
528 in dendritic cells, and (7) CD11c in dendritic cells, we employed a Gene Ontology analysis [48]
529 which identify top 10 pathways based on GO gene sets with FDR q-value < 0.05 as significant
530 (Table S4).

531

532 **Seurat anchor-transfer analysis**

533 We compared cTP-net with an anchor-based transfer learning method developed in Seurat v3
534 [20]. For Seurat v3, RNA count data are normalized by LogNormalization method, while surface
535 protein counts are normalized by centered log-ratio (CLR) method. In validation test setting, we

bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

536 used the same cells for training and testing as in cTP-net so as to be directly comparable to
537 cTP-net. For out-of-cell type prediction, default parameters did not work for several cell types in
538 anchor-transfer step, because, for those cell types, there are few anchors shared between the
539 training and testing sets. To overcome this, we reduced the number of anchors iteratively until
540 the function ran successfully.

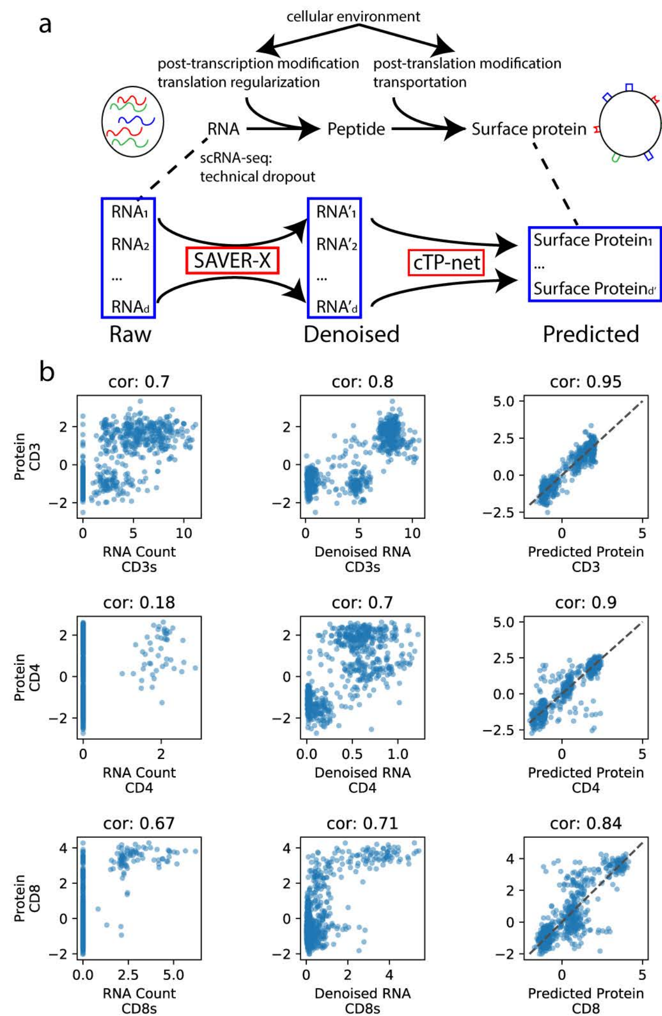
541

542 **HCA data analysis**

543 *HCA RNA-seq transcriptome data analysis.* HCA RNA-seq data sets are processed as
544 discussed above, resulting in log-normalized denoised values. We applied default pipeline of
545 Seurat and generated t-SNE plot for both data sets (Figure S7). Cells are clearly clustered by
546 individuals, indicating strong batch effects. As a result, the following analysis was performed on
547 cells of each individual. Major cell types were determined by known gene markers.

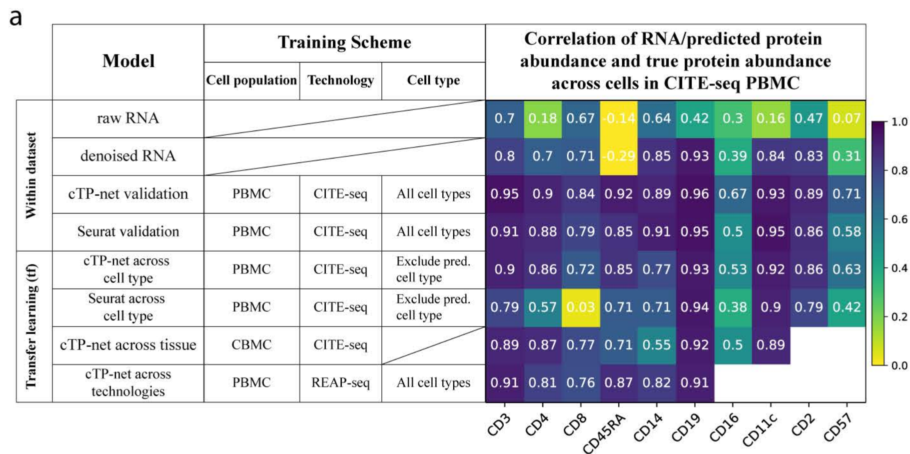
548 *Surface protein prediction by cTP-net.* From the log-normalized denoised expression value, we
549 predict the surface protein abundance with cTP-net model trained jointly on CITE-seq PBMC
550 and CBMC data sets. We embedded 12 surface protein abundance across 16 individuals on t-
551 SNE plot, showing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

bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



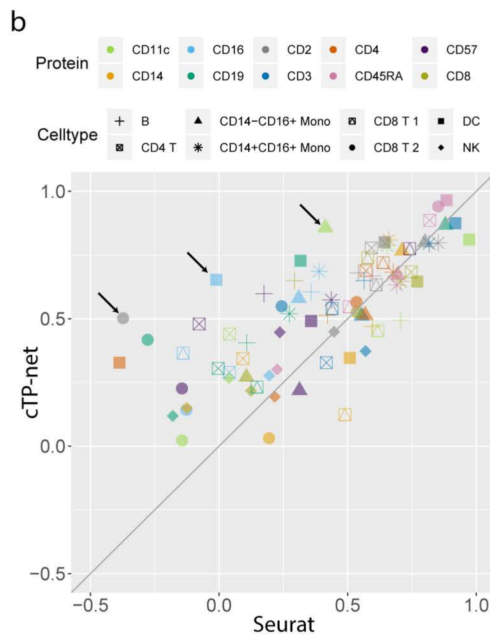
bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.

a

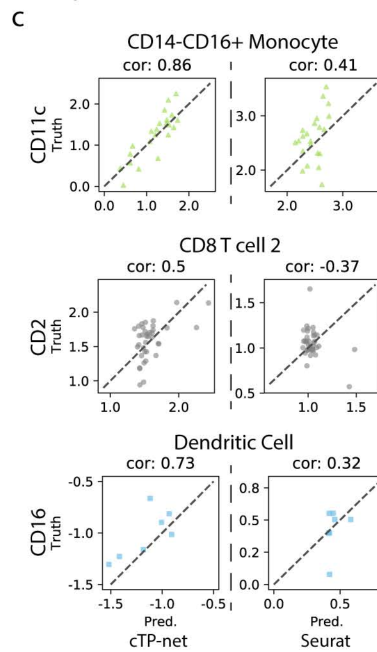


	Model	Training Scheme: Cell population	Training Scheme: Technology	Training Scheme: Cell type	CD3	CD4	CD8	CD45RA	CD14	CD19	CD16	CD11c	CD2	CD57
Within dataset	raw RNA				0.7	0.18	0.67	-0.14	0.64	0.42	0.3	0.16	0.47	0.07
Within dataset	denoised RNA				0.8	0.7	0.71	-0.29	0.85	0.93	0.39	0.84	0.83	0.31
Within dataset	cTP-net validation	PBMC	CITE-seq	All cell types	0.95	0.9	0.84	0.92	0.89	0.96	0.67	0.93	0.89	0.71
Within dataset	Seurat validation	PBMC	CITE-seq	All cell types	0.91	0.88	0.79	0.85	0.91	0.95	0.5	0.95	0.86	0.58
Transfer learning (tf)	cTP-net across cell type	PBMC	CITE-seq	Exclude pred. cell type	0.9	0.86	0.72	0.85	0.77	0.93	0.53	0.92	0.86	0.63
Transfer learning (tf)	Seurat across cell type	PBMC	CITE-seq	Exclude pred. cell type	0.79	0.57	0.03	0.71	0.71	0.94	0.38	0.9	0.79	0.42
Transfer learning (tf)	cTP-net across tissue	CBMC	CITE-seq		0.89	0.87	0.77	0.71	0.55	0.92	0.5	0.89		
Transfer learning (tf)	cTP-net across technologies	PBMC	REAP-seq	All cell types	0.91	0.81	0.76	0.87	0.82	0.91				

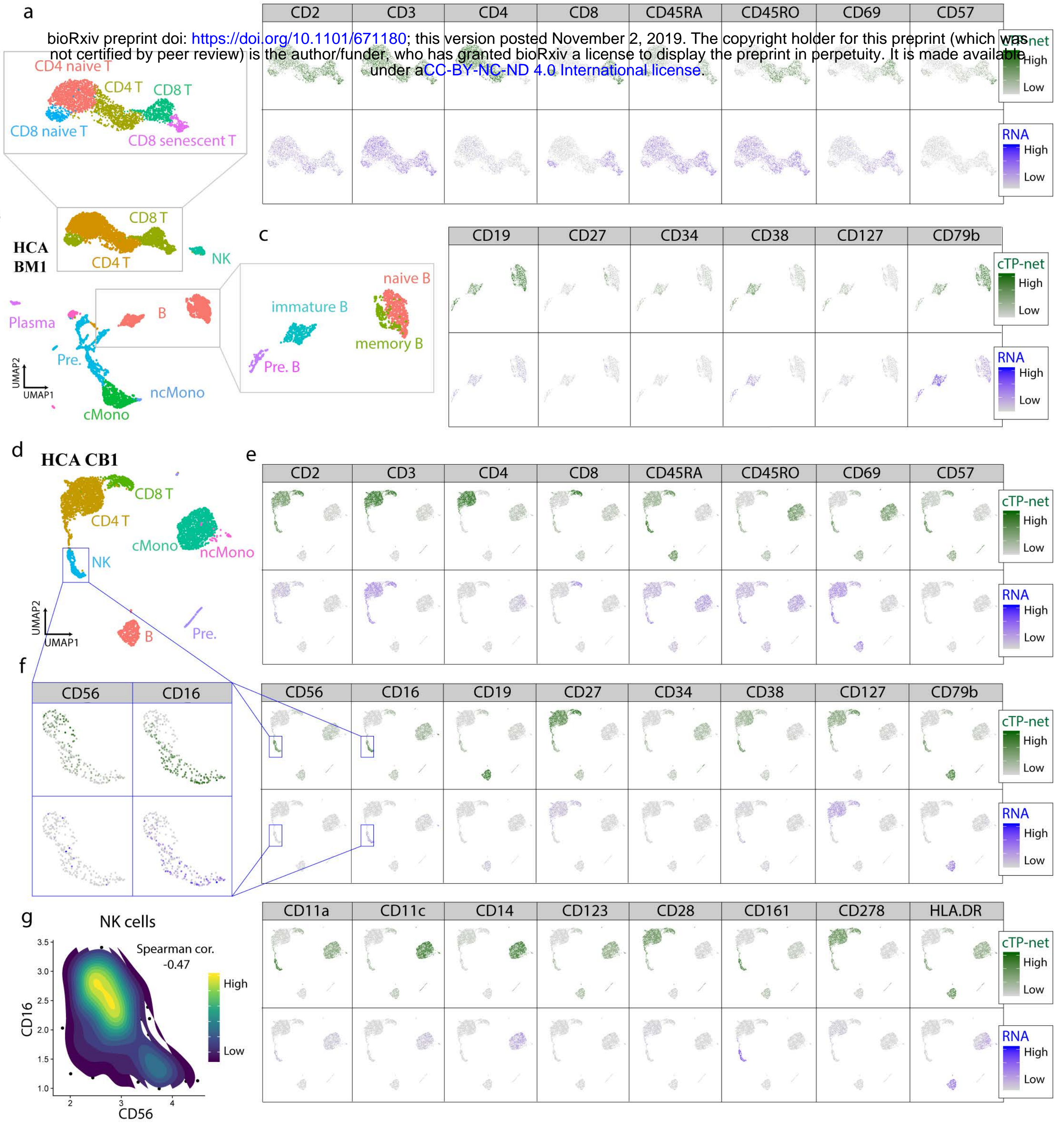
b



c



bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.



bioRxiv preprint doi: https://doi.org/10.1101/671180; this version posted November 2, 2019. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-NC-ND 4.0 International license.