

Supplementary Information

A multi-use deep learning method for CITE-seq and single-cell RNA-seq data integration with cell surface protein prediction and imputation

Justin Lakkis*, Amelia Schroeder, Kenong Su, Michelle Lee, Alexander C. Bashore, Muredach P. Reilly,
Mingyao Li*

***Correspondence:**

Justin Lakkis (jlakks@gmail.com)

Mingyao Li (mingyao@penncellmedicine.upenn.edu)

Supplementary Table 1. Datasets analyzed in this paper.

Data	Data Source	Number of Cells	Number of Genes	Number of proteins
MALT	10x Genomics https://www.10xgenomics.com/resources/datasets/10-k-cells-from-a-malt-tumor-gene-expression-and-cell-surface-protein-3-standard-3-0-0	8,412	33,538	17
PBMC	Hao et al. (2021) https://atlas.fredhutch.org/data/nygc/multimodal/pbmc_multimodal.h5seurat	161,764	20,729	224
Monocyte	Generated ourselves. Data will be publicly available after the paper is accepted for publication.	37,112	22,060	283
H1N1	Kotliarov et al. (2020) https://nih.figshare.com/articles/dataset/CITE-seq_protein-mRNA_single_cell_data_from_high_and_low_vaccine_responders_to_reproduce_Figs_4-6_and_associated_Extended_Data_Figs_/11349761?file=20706645	53,201	32,738	87
COVID (Haniffa)	Stephenson et al. (2021) https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-10026	647,366	24,737	192
COVID (Sanger)	Chan Zuckerberg Initiative Single-Cell COVID-19 Consortia (2020) https://covid19.cog.sanger.ac.uk/submissions/release2/vento_pbmc_processed.h5ad	240,627	33,567	192

Supplementary Note 1: Early stopping

Let ES_{max} denote the patience parameter for early stopping and LR_{max} denote the patience parameter for learning rate decay. Let $count$ be a counter which indicates the number of epochs that have elapsed since the validation loss decreased from its running minimum $best_{loss}$. After completing an epoch and computing the validation loss val_{loss} , check if $val_{loss} * 1.005 < best_{loss}$. If so, set $count = 0$ and update: $best_{loss} \leftarrow val_{loss}$. Otherwise, increment $count$ by 1. If $(count + 1) \bmod LR_{max}$ is equal to 0, decay the learning rate lr by a factor d . That is, make the update $lr \leftarrow lr \times d$. If $count$ equals ES_{max} , then end training as the validation loss has failed to decrease below its running minimum for ES_{max} consecutive epochs.

Supplementary Note 2: Software packages

We used totalVI via the scvi-tools package (<https://scvi-tools.org>). We specifically used version 0.10.0 of scvi-tools and version 4.1.0 of Seurat (<https://satijalab.org/seurat>). The sciPENN package can be found online on github (<https://github.com/jlakkis/sciPENN>).

All analyses can be reproduced using this repository (https://github.com/jlakkis/sciPENN_codes). All packages, including sciPENN, can be installed following the instructions in that repository.