A Joint Model of RNA Expression and Surface Protein Abundance in Single Cells

PREPRINT

Adam Gayoso[1], Romain Lopez[2], Zoë Steier[3], Jeffrey Regier[4], Aaron Streets[1, 3, 5, †], and Nir Yosef[1, 2, 5, 6, †]

[1] Center for Computational Biology, University of California, Berkeley

[2] Department of Electrical Engineering and Computer Sciences, University of California, Berkeley

[3] Department of Bioengineering, University of California, Berkeley

[4] Department of Statistics, University of Michigan, Ann Arbor

[5] Chan Zuckerberg Biohub, San Francisco, California

[6] Ragon Institute of MGH, MIT and Harvard

† Corresponding author: {astreets, niryosef}@berkeley.edu

October 4, 2019

1 Introduction

Single-cell RNA sequencing (scRNA-seq) enables comprehensive quantitative characterization of a cell's mRNA profile and has resulted in novel findings about the molecular circuitry of cell populations [1, 2]. Extending scRNA-seq, cellular indexing of transcriptomes and epitopes by sequencing (CITE-seq) simultaneously measures the abundance of selected proteins on the cell surface with results comparable to gold-standard flow cytometry [3]. As surface proteins are routinely used as a measure of cell phenotype, CITE-seq provides an exciting new opportunity for enhancing the quality of data interpretation [3, 4, 5], especially as the number of assayed proteins per experiment grows (with over 300 barcoded-antibodies commercially available [6]). Recent approaches to CITE-seq data analysis consist of deriving clusters based on the mRNA data and mapping protein values to these clusters for the task of cell-type labeling. However, this strategy assumes that cell-cell similarities depend only on mRNA, and neglects the distinctive information contained in the protein measurements. We aim to combine both measurements into one representation of cell state, while addressing the unique technical biases of each modality.

A reasonable modeling assumption is that both mRNA and protein counts are generated from a low-dimensional manifold of cellular states [1]. While a flourishing body of research has focused on using generative models to learn a biologically meaningful low-dimensional representation of cells in scRNA-seq datasets [7, 8, 9, 10], no proposed method can additionally handle the complexities of the protein counts. CITE-seq protein counts are overdispersed like

mRNA counts, but do not suffer from limited capture efficiency. Instead, the protein counts are obscured by a non-negligible background signal, which may arise due to non-specific binding of antibody probes and/or ambient antibodies. As a result, the distribution of protein counts across cells is often bimodal with a background and foreground component. A natural way to preprocess the protein counts is to fit a mixture model to each protein globally and replace a count with its probability of being generated from the larger component [11, 12]. However, there is no basis for assuming global bimodality of protein counts since a dataset typically comprises heterogeneous populations of cells each with distinct surface proteomes. Consequently, using the same signal-noise decision boundary for all cells can be inappropriate.

Here we propose Total Variational Inference (totalVI), a coupled generative model and inference procedure for CITE-seq data, which addresses these issues. In totalVI, both mRNA and protein counts of a cell are assumed to be random variables generated from a low-dimensional latent variable that represents the underlying biological state of a cell and contains information from both domains. Such a framework enables end-to-end analysis of this data – a joint batch-corrected latent representation (for stratifying cells into types), denoised data in both domains, and differential expression of genes and proteins. totalVI leverages advances in stochastic optimization and easily scales to millions of cells.

2 The totalVI probabilistic model

A CITE-seq experiment produces two vectors for a cell n, x_n and y_n, where x_{nr} is the number of mRNA molecules detected for gene r and y_{nt} is the number of cell surface molecules detected for protein t. Furthermore, a dataset has R genes and T assayed proteins. Let s_n be the batch cell n was processed in (one-hot encoded) with a total of B batches.

Let z_n be a latent variable describing the biological state of cell n. Given z_n and a cell-specific scaling factor ℓ_n representing technical factors like the mRNA sequencing depth, x_{nr} follows a negative binomial distribution with gene-specific inverse dispersion and with the prior for ℓ_n set as in [7]. Let μ_{nt} be a latent variable representing the mean of the background distribution for the protein counts, sampled from a protein-batch-specific prior. Given z_n and μ_{nt}, we model y_{nt} as a negative binomial mixture to capture observed protein counts arising from the background or foreground. The mean parameters of the mixture components are structured such that the foreground mean is strictly larger than the background mean, which also identifies the mixture as the inverse dispersion parameter is shared between components.

The full generative process is outlined in Algorithm 1. The prior parameters for μ_{nt} are learned in a variational Bayesian inference fashion. The neural networks are dense with one hidden layer, ReLU activations, batch normalization and dropout. Notably, z_n follows a logistic normal distribution, meaning cells can be interpreted as having "membership" to dimensions of the latent space and that archetypal analysis can be performed (not shown) [13, 14].

3 Inference

We use variational inference to obtain the approximate posterior distribution

$$q_\eta(\mu_n \mid z_n, s_n) q_\eta(z_n \mid x_n, y_n, s_n) q_\eta(l_n \mid x_n, s_n),$$

where η is the set of parameters of an inference network – a neural network that takes a cell's combined expression as input and outputs the parameters of the approximate posterior. The variational distribution $q_\eta(\mu_n \mid z_n)$ has parameters specific to the cell n and not global parameters like the prior. We optimize the evidence lower bound (ELBO) [15] of $\log p_\nu(x_{1:N}, y_{1:N} \mid s_{1:N})$ with respect to the variational parameters η and model parameters ν using stochastic gradients [16]. To avoid inference over discrete random variables, we analytically integrate out v_{nt}, yielding $p_\nu(y_{nt}|z_n, \mu_{nt})$, which is a mixture of negative binomials. We use the Adam optimizer [17], along with deterministic warm-up, and a reduction of the learning rate upon plateau of the ELBO on a validation set.

Algorithm 1: The totalVI generative model. The negative binomial distribution is parameterized by its mean and inverse dispersion. Let ν be the set of model parameters described here.

Define: Neural networks $f(z_n, s_n; \Lambda) : \Delta^{K-1} \times \{0,1\}^B \to \Delta^{R-1}$,
$h(z_n, s_n; \Lambda) : \Delta^{K-1} \times \{0,1\}^B \to \mathbb{R}^T$

Require: Inverse dispersion parameters $\theta \in \mathbb{R}_+^R, \phi \in \mathbb{R}_+^T$. Neural network parameters Ξ, Ψ, Ω.

for *each cell* n **do**

- $z_n \sim \text{LogisticNormal}(0, I)$ — K-dimensional biological state variable
- $\rho_n = f(z_n, s_n; \Xi)$ — R-dimensional mRNA frequency
- $\alpha_n = \text{ReLU}(h(z_n, s_n; \Psi)) + 1$ — T-dimensional foreground increment protein scaling
- $\pi_n = \text{sigmoid}\,(h(z_n, s_n; \Omega))$ — T-dimensional mixture parameter
- $\ell_n \sim \text{LogNormal}(\ell_\mu, \text{diag}(\ell_{\sigma^2})I)$ — B-dimensional cell scaling factor for mRNA
- **for** *each gene* r **do**
  - $x_{nr} \sim \text{NegativeBinomial}\left(\rho_{nr} s_n^\top \ell_n, \theta_r\right)$
- **for** *each protein* t **do**
  - $\mu_{nt} \sim \text{LogNormal}(c_t^\top s_n, d_t^\top s_n)$ — Scalar background mean
  - $v_{nt} \sim \text{Bernoulli}(\pi_{nt})$ — Scalar foreground/background mixing variable
  - **if** $v_{nt} = 1$ **then**
    - $y_{nt} \sim \text{NegativeBinomial}\,(\mu_{nt}, \phi_t)$
  - **else**
    - $y_{nt} \sim \text{NegativeBinomial}\,(\mu_{nt}\alpha_{nt}, \phi_t)$

4 Performance benchmarks

We assess the performance of totalVI on two tasks: generalization to held-out data and posterior predictive checks (PPC) of coefficient of variation. We compare totalVI to factor analysis (FA) in which we either fit on the data that is logarithmized plus one (log), or log normalized data where the two modalities are independently normalized by their library size prior to log transformation (log rate). We also compare to the state-of-the-art scRNA-seq model, scVI [7], where we treat the protein data as additional features (genes) and use a negative binomial likelihood distribution for both mRNA and protein counts.

For each of these tasks, we use two datasets: (1) 7,225 peripheral blood mononuclear cells (PBMC10k) from 10X Genomics [18] and (2) 8,412 cells from a MALT tumor (MALT) [19]. We first remove genes that are expressed in fewer than 1% of cells and retain the top 5,000 genes as measured by variance across cells. Both datasets contain 14 proteins. We filter cells that were below the first percentile for protein UMI counts and had expressed fewer than 500 genes. For PBMC10k we filter doublets using DoubletDetection [20].

Generalization to held-out data We compare totalVI to scVI and the FA baselines using 20 latent dimensions for each method. For each model, we compute separately for protein and gene features the mean squared logarithmic error (MSLE) between the observed values and the mean of the posterior predictive distribution on a held-out subset representing 6% of the total dataset. For example,

$$\text{MSLE}_{\text{RNA}} = \frac{1}{NR} \sum_{n,r} \left(\log \frac{\hat{x}_{nr} + 1}{x_{nr} + 1} \right)^2$$

where $\hat{x}_{nr} = \mathbb{E}_{p_\nu(x^*_{nr} | x_n, y_n)}\left[x^*_{nr}\right]$, the mean of the posterior predictive, while $\text{MLSE}_{\text{Protein}}$ is computed using $y_{1:N}$ instead. Table 1 shows that totalVI has lower MSLE with respect to mRNA and proteins, which can be attributed to the superior noise model for proteins used in totalVI.

We also compute the held-out log-likelihood for totalVI and scVI. Table 2 shows that totalVI outperforms scVI on both datasets. As log-likelihoods for models with discrete and continuous likelihoods are not directly comparable, we computed the calibration error [21] in order to quantify the quality of each model's uncertainty estimates. totalVI and scVI have lower calibration error than FA models (results not shown).

	PBMC10k		MALT	
Model	Protein	RNA	Protein	RNA
FA (log)	0.935	0.167	0.938	0.169
FA (log rate)	0.860	0.125	0.870	0.121
scVI	0.739	0.105	0.474	0.103
totalVI	**0.599**	**0.103**	**0.431**	**0.098**

Table 1: Mean squared logarithmic error between observations and the mean of the posterior predictive distribution.

Model	PBMC10k	MALT
scVI	3349.44	3180.80
totalVI	**3337.89**	**3158.78**

Table 2: Negative log likelihood on held-out data.

	PBMC10k		MALT	
Model	Protein	RNA	Protein	RNA
FA (log)	1.515	2.119	8.075	2.647
FA (log rate)	0.691	1.146	1.646	1.466
scVI	0.389	0.082	0.725	0.139
totalVI	**0.164**	**0.023**	**0.620**	**0.066**

Table 3: CV PPC. Median absolute error.

Posterior predictive checks We perform a PPC [22] of the coefficient of variation (CV) for each protein and each gene. For each model, we sample the posterior predictive distribution 25 times, calculating the CV for each sample and for each feature. After averaging over samples, we separate the predicted CVs by feature type (mRNA or protein), resulting in two vectors: CV_{RNA} and $\text{CV}_{\text{Protein}}$. We report the median absolute error between the observed and the predicted CV_{RNA} and $\text{CV}_{\text{Protein}}$. Table 3 shows that totalVI outperforms other methods in both modalities indicating better model fit.

5 Protein background disentanglement

Disentangling background and foreground protein counts is crucial for mitigating spurious differential expression and cell-type labeling results. Previous work derives a linear cutoff on the number of counts for each protein based on spiked-in cells that do not express the proteins specifically recognized by the barcoded antibodies [3]. Others fit mixture models to each protein, which assumes that all cells are subject to the same background distribution [11, 12]. Our approach, which obviates the need for negative control cells or the assumption of a constant background distribution, models each $y_{nt}|z_n, \mu_{nt}$ as a negative binomial mixture, where the Bernoulli parameter $\pi_{nt}|z_n$ can be interpreted as the probability that a cell's protein count came from the background. Thus, decision boundaries are cell-protein-specific, taking into account the overall state of the cell (genes and proteins).

As an example, consider the CD16 protein in PBMC10k (Figure 1). We highlight a subset of cells that could be called background by a global mixture model but are predicted to be foreground by totalVI (blue: $\mathbb{E}_{q_\eta(z_n|x_n,y_n)}[\pi_{nt}] < 0.1$). We also highlight a subset of cells with similar magnitude of expression as the blue set but with higher predicted background probability (red: $\mathbb{E}_{q_\eta(z_n|x_n,y_n)}[\pi_{nt}] > 0.9$). Consistent with the model, we observe that the foreground-predicted subset (blue) corresponds to natural killer cells and CD16+ monocytes (both of which are known to express CD16), while the background subset of cells (red)

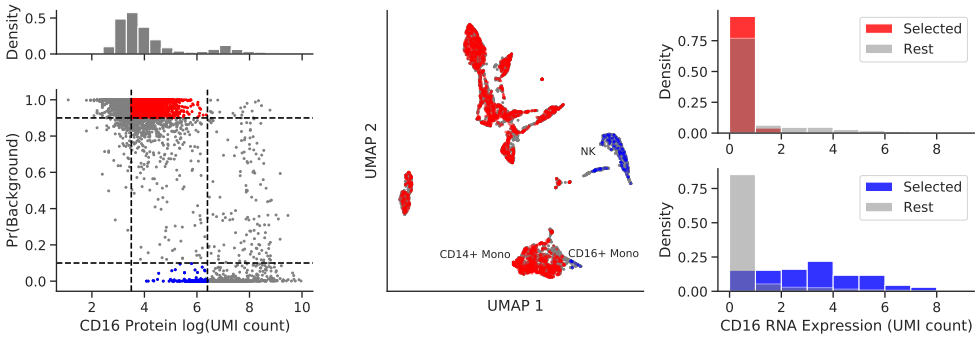

Figure 1: Investigation of totalVI background prediction for CD16 protein counts in PBMC10k.

correspond to the other PBMC cell types (which do not tend to express CD16; note that cell-types were determined by mRNA).

Figure 1 also shows that mRNA counts of the CD16 gene are high in the foreground and low in the background subset. Thus, totalVI makes a non-trivial prediction that goes beyond a simple cutoff – by leveraging information between all cells, genes, and proteins.

Despite using a two-component mixture density, totalVI can also disentangle the background of proteins that are trimodal globally. For example, it has been shown using flow cytometry that monocytes have fewer CD4 protein molecules on their surface relative to CD4+ T-cells [23] and that other PBMC types do not tend to express CD4 on their surface. As such, the distribution of the CD4 protein in PBMC10k is trimodal (lowest mode corresponding to background).

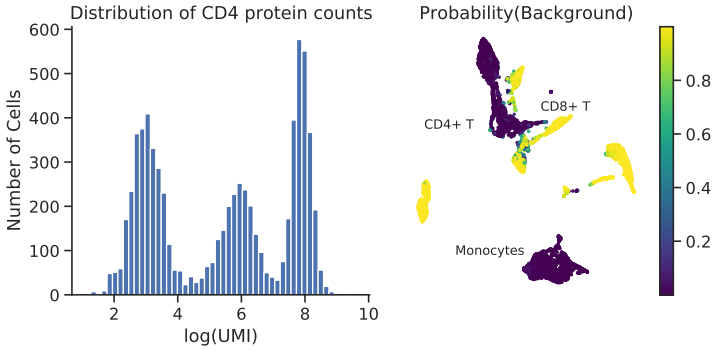

Figure 2: CD4 Protein disentanglement in PBMC10k. (Left) Distribution of log counts for CD4 protein. (Right) $\mathbb{E}_{q_\eta(z_n|x_n,y_n)}[\pi_{nt}]$ projected on UMAP.

Figure 2 shows the trimodal distribution of CD4 as well as the probability of background mapped to a UMAP [24] projection of $\mathbb{E}_{q_\eta(z_n|x_n,y_n)}[z_n]$. The cells we manually labeled (based on mRNA) as CD4+ T-cells and monocytes are indeed determined to have been mostly generated from the foreground component, while the remaining cells fall within the background part. This result is made possible due to the mixture being conditionally dependent on z_n, thus defining the two modes of the foreground-background dichotomy in a manner local to the latent space. These predictions will be critical for quantifying differential protein expression.

6 Data denoising

Another application of totalVI is data denoising, in which a denoised expression matrix (mRNA and protein) is produced that can then be used as input for other downstream tasks, like building a feature-wise correlation matrix to identify signaling and regulatory networks, or examining the dynamics of transcription and translation. The totalVI denoised expression matrix is constructed by first replacing x_{nr} with $\mathbb{E}_{q_\eta(z_n|x_n)}[\rho_{nr}]$ for all cells and all genes. For a protein count y_{nt}, we replace with the quantity $\mathbb{E}_{p_\nu(y^*_{nt}, v_{nt}=0, z_n|x_n, y_n)}[y^*_{nt}]$ which is interpreted as the expected protein count given it was generated by the foreground component of the mixture, adjusted for the probability it was generated by the foreground. It is then normalized so that the denoised protein expression for a cell n resides in the simplex. Thus, protein counts likely to have been generated from the background will have magnitude near zero in the denoised matrix.

We construct Spearman correlation matrices between all proteins and 500 genes chosen randomly except for the inclusion of genes that encode for the assayed proteins. The matrices are derived using (1) denoised expression, (2) raw counts, (3) posterior predictive counts.

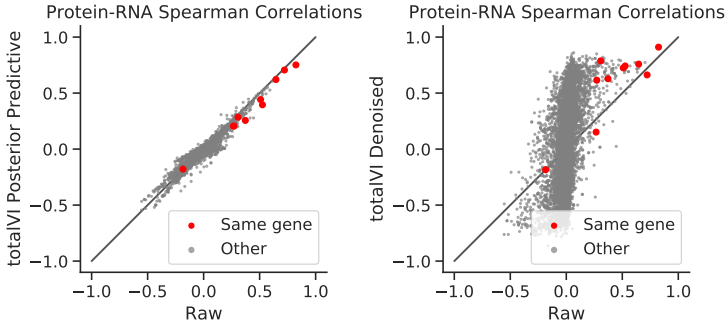

Figure 3: totalVI posterior predictive and denoised Spearman correlations versus raw correlations in PBMC10k.

In Figure 3 we plot the posterior predictive correlations relative to the raw correlations as well as the denoised correlations relative to the raw correlations. Correlations of features derived from the same gene are in red while all others are in grey. We emphasize that totalVI has no prior knowledge of RNA-protein translation relationships. We see that the denoised correlations are more extreme than their raw counterparts, however posterior predictive correlations largely match the original raw correlations, indicating that extreme denoised correlations are not symptomatic of fitting a low-dimensional model and that totalVI is instead likely to restore biological correlations.

7 Dataset harmonization

We demonstrate how totalVI can be used to generate a batch-corrected joint latent representation by harmonizing two PBMC datasets (PBMC10k and another dataset of 5k PBMCs from 10X Genomics [25] with genes and proteins subsetted to match PBMC10k; results in Figure 4). totalVI has a unique advantage over popular methods based on mutual nearest neighbors [4, 26], as no similarity metric between cells is necessary, which may be biased toward one modality. Instead, independence between z_n and s_n is a byproduct of the invariance of the prior on z_n to the batch. Another benefit of our method is the ability to marginalize over batch in order to generate batch-free denoised values, or perform differential expression over batches.

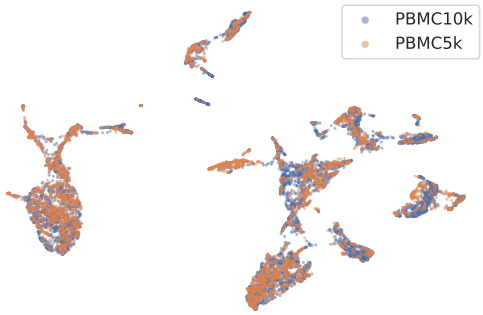

Figure 4: totalVI batch-corrected joint latent representation visualized with UMAP.

Code availability

The implementation to reproduce the experiments of this paper is available at https://github.com/adamgayoso/totalVI_reproducibility. The reference implementation of totalVI is available at https://github.com/YosefLab/scVI. Datasets are publicly available from 10X Genomics.

Acknowledgements

AG is supported by NIH Training Grant 5T32HG000047-19. ZS is supported by the NSF GRFP. This work is supported in part by NIH Grant R35GM124916. AS and NY are Chan Zuckerberg Biohub investigators.

References

[1] Allon Wagner, Aviv Regev, and Nir Yosef. Revealing the vectors of cellular identity with single-cell genomics. *Nature biotechnology*, 2016.

[2] Amos Tanay and Aviv Regev. Scaling single-cell genomics from phenomenology to mechanism. *Nature*, 2017.

[3] Marlon Stoeckius, Christoph Hafemeister, William Stephenson, Brian Houck-Loomis, Pratip K Chattopadhyay, Harold Swerdlow, Rahul Satija, and Peter Smibert. Simultaneous epitope and transcriptome measurement in single cells. *Nature methods*, 2017.

[4] Tim Stuart, Andrew Butler, Paul Hoffman, Christoph Hafemeister, Efthymia Papalexi, William M Mauck III, Yuhan Hao, Marlon Stoeckius, Peter Smibert, and Rahul Satija. Comprehensive integration of single-cell data. *Cell*, 2019.

[5] Jeffrey M Granja, Sandy Klemm, Lisa M McGinnis, Arwa S Kathiria, Anja Mezger, Benjamin Parks, Eric Gars, Michaela Liedtke, Grace XY Zheng, Howard Y Chang, et al. A single cell framework for multi-omic analysis of disease identifies malignant regulatory signatures in mixed phenotype acute leukemia. *bioRxiv*, 2019.

[6] Antibody oligonucleotide conjugation services: TotalSeq and CITE-seq focus.

[7] Romain Lopez, Jeffrey Regier, Michael B. Cole, Michael I. Jordan, and Nir Yosef. Deep generative modeling for single-cell transcriptomics. *Nature Methods*, 2018.

[8] Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell rna-seq data. *Nature communications*, 2018.

[9] Hanna Mendes Levitin, Jinzhou Yuan, Yim Ling Cheng, Francisco JR Ruiz, Erin C Bush, Jeffrey N Bruce, Peter Canoll, Antonio Iavarone, Anna Lasorella, David M Blei,

et al. De novo gene signature identification from single-cell rna-seq with hierarchical poisson factorization. *Molecular systems biology*, 2019.

[10] Sandhya Prabhakaran, Elham Azizi, Ambrose Carr, and Dana Pe'er. Dirichlet process mixture model for correcting technical variation in single-cell gene expression data. In *International Conference on Machine Learning*, 2016.

[11] Trung Ngo Trong, Roger Kramer, Juha Mehtonen, Gerardo González, Ville Hautamäki, and Merja Heinäniemi. Sisua: Semi-supervised generative autoencoder for single cell data. *ICML Workshop in Computational Biology*, 2019.

[12] Kiya W Govek, Emma C Troisi, Steven Woodhouse, and Pablo G Camara. Single-cell transcriptomic analysis of mihc images via antigen mapping. *bioRxiv*, 2019.

[13] Adele Cutler and Leo Breiman. Archetypal analysis. *Technometrics*, 1994.

[14] David Van Dijk, Roshan Sharma, Juozas Nainys, Kristina Yim, Pooja Kathail, Ambrose J Carr, Cassandra Burdziak, Kevin R Moon, Christine L Chaffer, Diwakar Pattabiraman, et al. Recovering gene interactions from single-cell data using data diffusion. *Cell*, 2018.

[15] David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational Inference: A Review for Statisticians. *Journal of the American Statistical Association*, 2017.

[16] Diederik P Kingma and Max Welling. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations*, 2014.

[17] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015.

[18] 10X Genomics. 10k PBMCs from a healthy donor - gene expression and cell surface protein. 2018.

[19] 10X Genomics. 10k cells from a MALT tumor - gene expression and cell surface protein. 2018.

[20] Adam Gayoso and Jonathan Shor. GitHub: DoubletDetection, 2019.

[21] Volodymyr Kuleshov, Nathan Fenner, and Stefano Ermon. Accurate uncertainties for deep learning using calibrated regression. In *Proceedings of the 35th International Conference on Machine Learning*, 2018.

[22] Andrew Gelman, Xiao-Li Meng, and Hal Stern. Posterior predictive assessment of model fitness via realized discrepancies. *Statistica sinica*, 1996.

[23] Lionel G Filion, Carlos A Izaguirre, Gary E Garber, Lothar Huebsh, and Maung T Aye. Detection of surface and cytoplasmic cd4 on blood monocytes from normal and hiv-1 infected individuals. *Journal of immunological methods*, 1990.

[24] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv*, 2018.

[25] 10X Genomics. 5k peripheral blood mononuclear cells (PBMCs) from a healthy donor with cell surface proteins (v3 chemistry). 2019.

[26] Laleh Haghverdi, Aaron TL Lun, Michael D Morgan, and John C Marioni. Batch effects in single-cell rna-sequencing data are corrected by matching mutual nearest neighbors. *Nature biotechnology*, 2018.