nature portfolio

https://doi.org/10.1038/s41592-020-01050-x

Supplementary information

Joint probabilistic modeling of single-cell multi-omic data with totalVI

In the format provided by the authors and unedited

Joint probabilistic modeling of single-cell multi-omic data with totalVI

SUPPLEMENTARY INFORMATION

Adam Gayoso [1,*], Zoë Steier [2,*], Romain Lopez [3], Jeffrey Regier [4], Kristopher L Nazor [5], Aaron Streets [1,2,6,†], and Nir Yosef [1,3,6,7,†]

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

[2] Department of Bioengineering,
University of California, Berkeley,
Berkeley, CA, USA

[3] Department of Electrical Engineering and Computer Sciences,
University of California, Berkeley,
Berkeley, CA, USA

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

[5] BioLegend, Inc.,
San Diego, CA, USA

[6] Chan Zuckerberg Biohub,
San Francisco, CA, USA

[7] Ragon Institute of MGH, MIT and Harvard

* These authors contributed equally.

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

Contents

Supplementary Tables

Dataset name	Antibody panel	Day	Mouse	Tissue	Cells (captured)	Cells (post-filtering)
SLN111-D1	111	1	A	Spleen; Lymph Node	11,160	9,264
SLN111-D2	111	2	B	Spleen; Lymph Node	9,017	7,564
SLN208-D1	208	1	A	Spleen; Lymph Node	10,777	8,715
SLN208-D2	208	2	B	Spleen; Lymph Node	8,921	7,105

Supplementary Table 1: Summary of spleen and lymph node datasets. Each dataset was processed in a separate 10x lane. Each day indicates a 10x run. Cells captured is the number of cells reported by Cell Ranger.

Cell type	Protein	totalVI ROC AUC	GMM ROC AUC
B cells	CD19	**0.9998**	0.9997
B cells	CD45R-B220	**0.9999**	0.9974
B cells	CD20	**0.9995**	0.8600
B cells	I-A-I-E	**0.9941**	0.9725
T cells	CD5	**0.9998**	0.9974
T cells	TCRbchain	**0.9999**	0.9976
T cells	CD90.2	**0.9999**	0.9986
T cells	CD28	**0.9999**	0.8328
CD4 T cells	CD4	**0.9999**	0.9969
CD8 T cells	CD8a	**0.9999**	0.9985
CD8 T cells	CD8b	**0.9998**	0.9980

Supplementary Table 2: Classification of cell types by marker proteins (SLN111-D1 dataset.) Performance of totalVI and a Gaussian mixture model (GMM) fit on all cells for each protein of the SLN111-D1 dataset. Area under the receiver operating characteristic curve (ROC AUC score) was calculated using as input either the totalVI foreground probability or GMM foreground probability where the indicated cell type was the positive population out of all B and T cells. AUC results were truncated at four decimal places. Bolded ROC AUC scores indicate higher values (better performance). Highlighted in blue are two proteins for which totalVI noticeably outperformed the GMM.

Cell type	Protein	Isotype Control	totalVI ROC AUC	GMM ROC AUC	GMM norm1 ROC AUC	GMM norm2 ROC AUC
B cells	CD19	IsotypeCtrlRatIgG2a_k	**0.9999**	0.9996	0.9988	0.9988
B cells	CD45R-B220	IsotypeCtrlRatIgG2a_k	**0.9999**	0.9973	0.9959	0.9961
B cells	CD20	IsotypeCtrlRatIgG2b_k	**0.9999**	0.7197	0.7528	0.7494
B cells	I-A-I-E	IsotypeCtrlRatIgG2b_k	**0.9984**	0.9783	0.9695	0.9695
T cells	CD5	IsotypeCtrlRatIgG2a_k	**0.9998**	0.9966	0.9938	0.9938
T cells	TCRbchain	IsotypeCtrlArm.HamsterIgG	**0.9999**	0.9943	0.7725	0.9762
T cells	CD90.2	IsotypeCtrlRatIgG2b_k	**0.9999**	0.9988	0.8316	0.9984
T cells	CD28	NA	**0.9997**	0.7805	0.7805	0.7805
CD4 T cells	CD4	IsotypeCtrlRatIgG2a_k	**0.9999**	0.9966	0.9966	0.9966
CD8 T cells	CD8a	IsotypeCtrlRatIgG2a_k	**0.9999**	0.9991	0.7952	0.9989
CD8 T cells	CD8b	IsotypeCtrlRatIgG2a_k	**0.9999**	0.9993	0.8606	0.9989

Supplementary Table 3: Classification of cell types by marker proteins (SLN208-D1 dataset). Performance of totalVI and a GMM fit on all cells for each protein of the SLN208-D1 dataset as in Supplementary Table 2. GMM norm1 indicates that data were normalized using an isotype control as in Cumulus [1] prior to fitting the GMM; GMM norm2 indicates that data were normalized using a modification to the Cumulus method (Methods). Bolded ROC AUC scores indicate highest value for each protein. Notable result is highlighted in blue.

Cell type annotation	Selected markers	Selected references
Activated CD4 T cells	*Itm2a, Cd69*	[2]
B1 B cells	*Bhlhe41*, CD43, CD19	[3, 4]
CD122+ CD8 T cells	CD122, CD62L, CD183(CXCR3), CD8a	[5, 6]
CD4 T cells	CD4	[7]
CD8 T cells	CD8a, CD8b	[7]
cDC1s	*Clec9a, Cd8a, Xcr1*, CD11c	[8, 9]
cDC2s	*Itgax, Cd4*, CD11c	[8, 9]
Cycling B/T cells	*Birc5, Top2a, Mki67*	[10]
Erythrocytes	*Hbb-bs, Hbb-bt*	[11]
GD T cells	*Cd3e, Tcrg-c1, Tcrg-c2, Maf, Il17re*	[12]
ICOS-high Tregs	*Foxp3*, CD4, ICOS	[13, 14]
Ifit3-high B cells	*Ifit3*, CD19	
Ifit3-high CD4 T cells	*Ifit3*, CD4	
Ifit3-high CD8 T cells	*Ifit3*, CD8a	
Ly6-high monocytes	*Ly6c2, Fn1, F13a1*	[15]
Ly6-low monocytes	*Cd36, Cd300e, Fabp4*	[16]
Mature B cells	IgD, CD23, CD19	[17]
Migratory DCs	*Slco5a1, Anxa3, Nudt17, Adcy6, Cacnb3*	[18]
MZ B cells	CD21, CD19	[17]
MZ/Marco-high macrophages	*Cd209b, Marco*	[19]
Neutrophils	*S100a8*	[20]
NK cells	NK-1.1, *Gzma, Ncr1*	[7, 21]
NKT cells	NK-1.1, *Cd3e, Ccl5, Klrd1*	[7, 22]
pDCs	*Siglech, Irf8, Runx2*, CD11c	[8, 23, 24]
Plasma B cells	*Jchain*	[25]
Red-pulp macrophages	F4-80, *C1qa, C1qb, Hmox1, Vcam1*	[26]
Transitional B cells	CD93, CD24, CD19	[27, 28]
Tregs	*Foxp3*, CD4, CD357(GITR)	[29]

Supplementary Table 4: Annotated cell types and selected markers in the spleen and lymph node datasets. cDC1: Conventional dendritic cell 1. cDC2: Conventional dendritic cell 2. GD: Gamma/delta. MZ: Marginal zone. NK: Natural killer. NKT: Natural killer T. pDC: Plasmacytoid dendritic cell. Treg: Regulatory CD4 T cell.

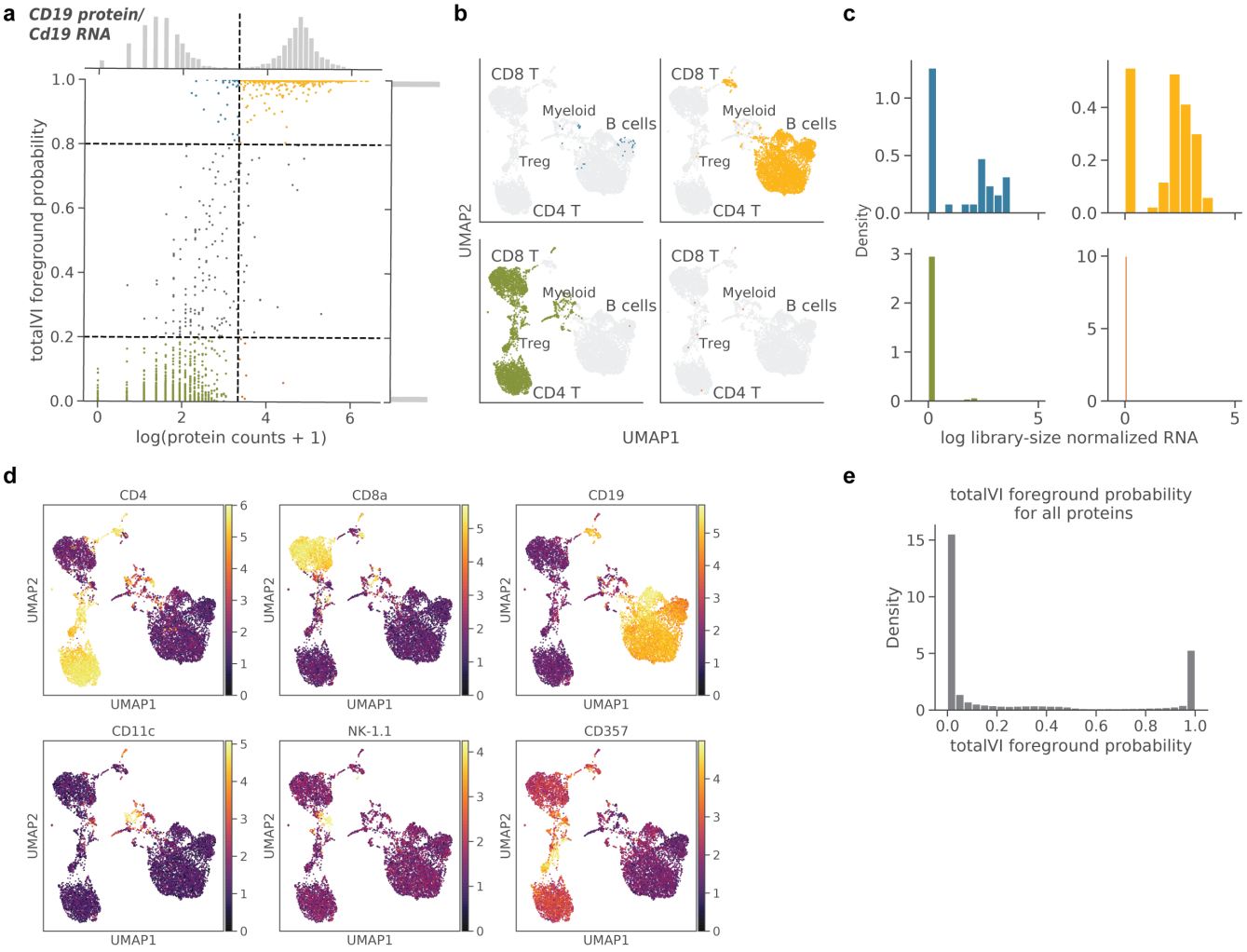
Name	RNA reads	Protein reads	RNA UMI	Protein UMI
SLN111-D1	34,717	4,733	4,392	2,785
SLN111-D2	45,765	6,542	2,121	3,419
SLN208-D1	33,569	5,513	4,561	2,956
SLN208-D2	43,821	3,961	2,102	2,261

Supplementary Table 5: Sequencing statistics for spleen and lymph node datasets. Sequencing statistics calculated per 10x lane by Cell Ranger. RNA reads: mean reads per cell from RNA. Protein reads: mean reads per cell from antibody barcodes. RNA UMI: median UMI counts per cell from RNA. Protein UMI: median UMI counts per cell from antibody barcodes.

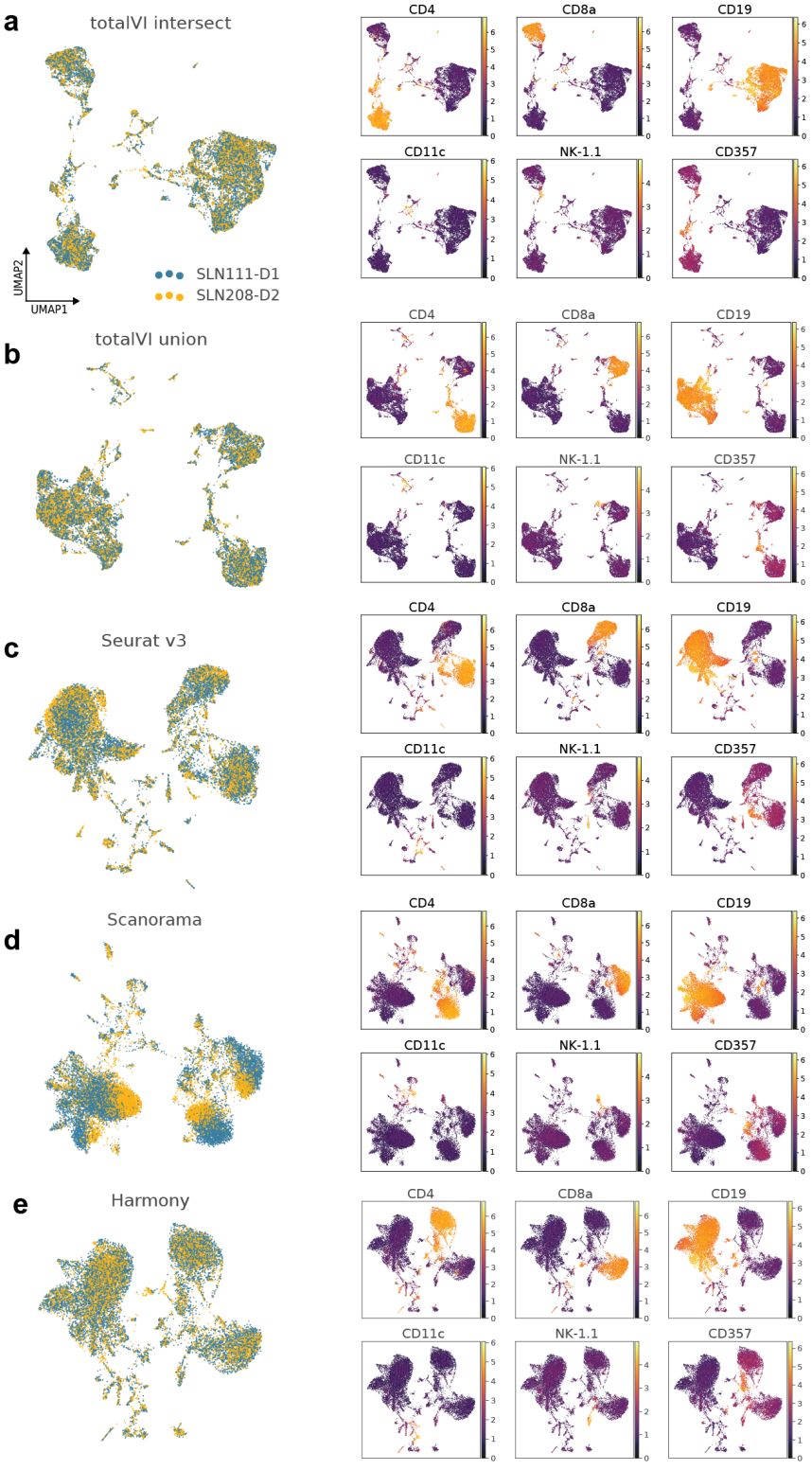
Dataset	No. cells	Pct. Mito	Protein Lib Size Range	No. Genes Expr.	RNA lib size
PBMCK10k	6,855	$< 10\%$	[400, 20,000]	$<$ 4,500	$<$ 20,000
PBMC5k	3,994	$< 20\%$	[400, 20,000]	$<$ 4,500	$<$ 20,000
MALT	6,838	$< 15\%$	[400, 20,000]	$<$ 5,000	$<$ 30,000

Supplementary Table 6: Summary of filtering parameters for publicly available datasets. Ranges indicate criteria for retained cells.

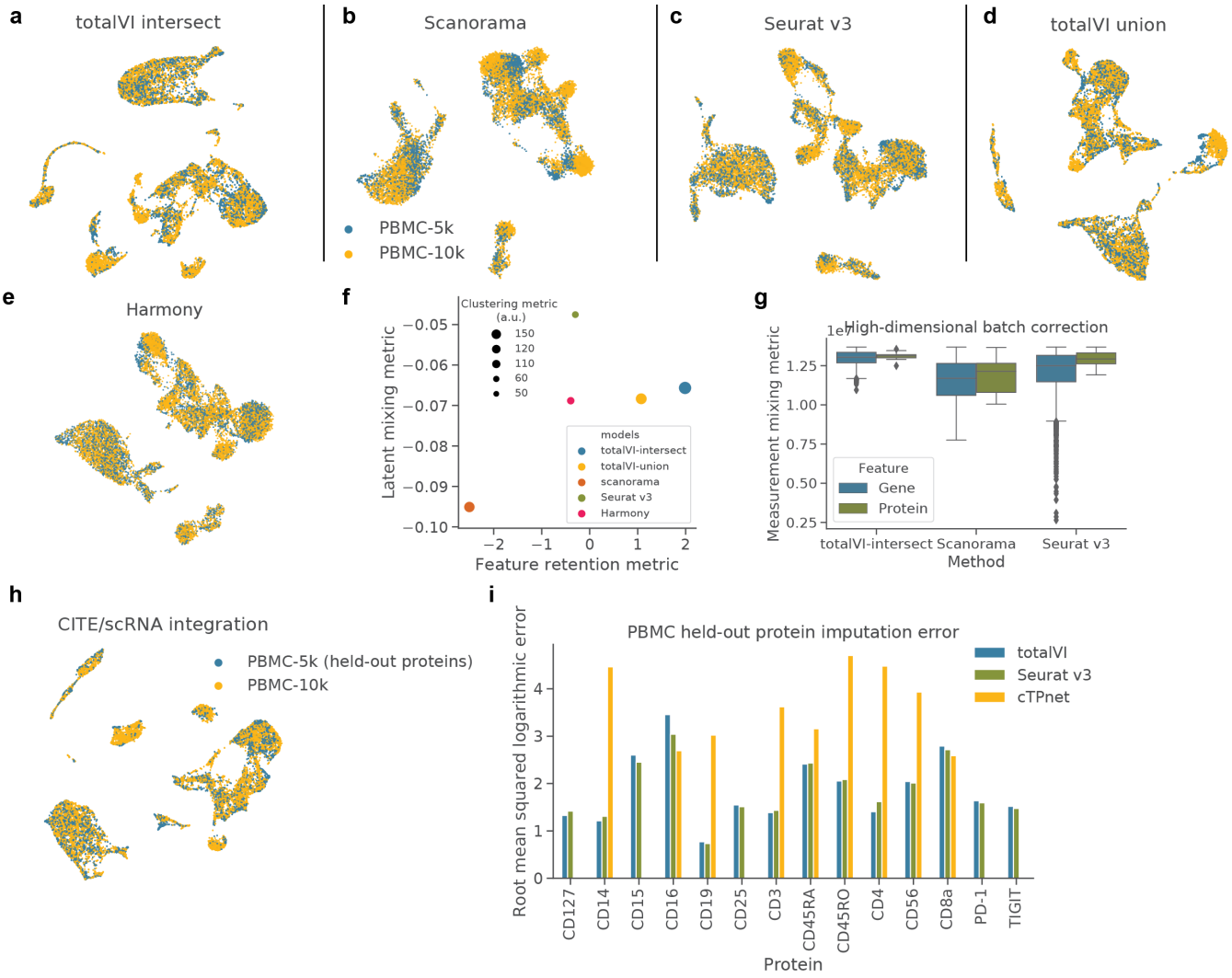
Supplementary Figures



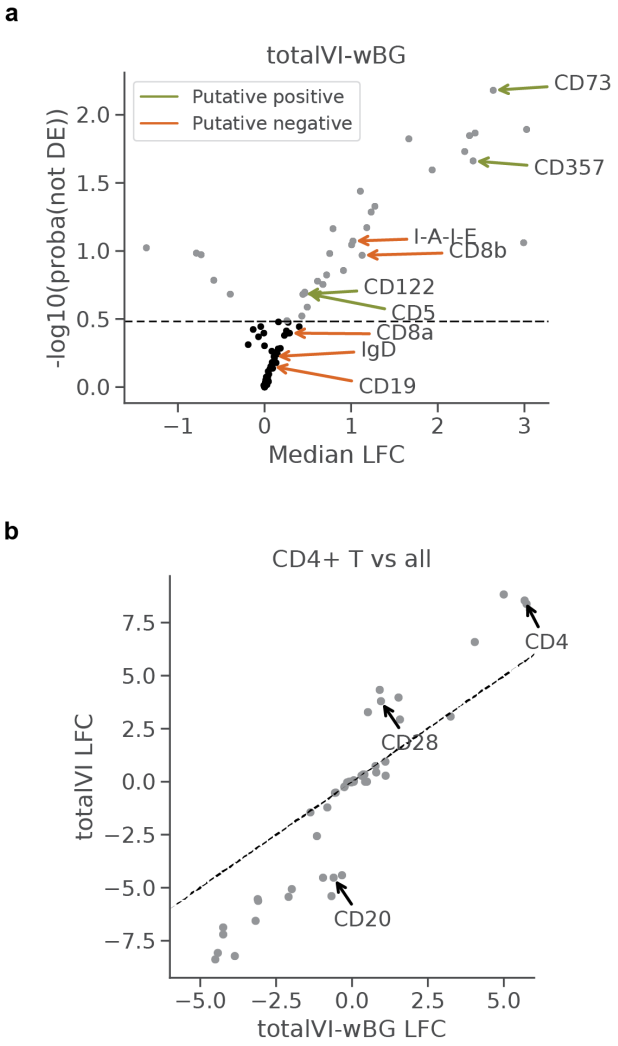
Supplementary Figure 1: totalVI decouples protein foreground and background. totalVI was applied to the SLN111-D1 dataset. **a-c**, CD19 protein (encoded by *Cd19* RNA). **(a)** totalVI foreground probability vs $\log(\text{protein counts} + 1)$. Vertical line denotes protein foreground/background cutoff determined by a GMM. Horizontal lines denote foreground probability of 0.2 and 0.8. Cells with foreground probability greater than 0.8 or less than 0.2 are colored by quadrant, while the remaining cells are gray. **(b)** UMAP plots of the totalVI latent space. Each quadrant contains cells from the corresponding quadrant of (a) in color with the remaining cells in gray. **(c)** RNA expression (log library-size normalized; Methods) for cells colored in (a). **d**, UMAP plots of the totalVI latent space colored by ($\log(\text{counts} + 1)$) of cell type marker proteins (Supplementary Table 4). **e**, totalVI foreground probability for all proteins across all cells in the SLN111-D1 dataset.



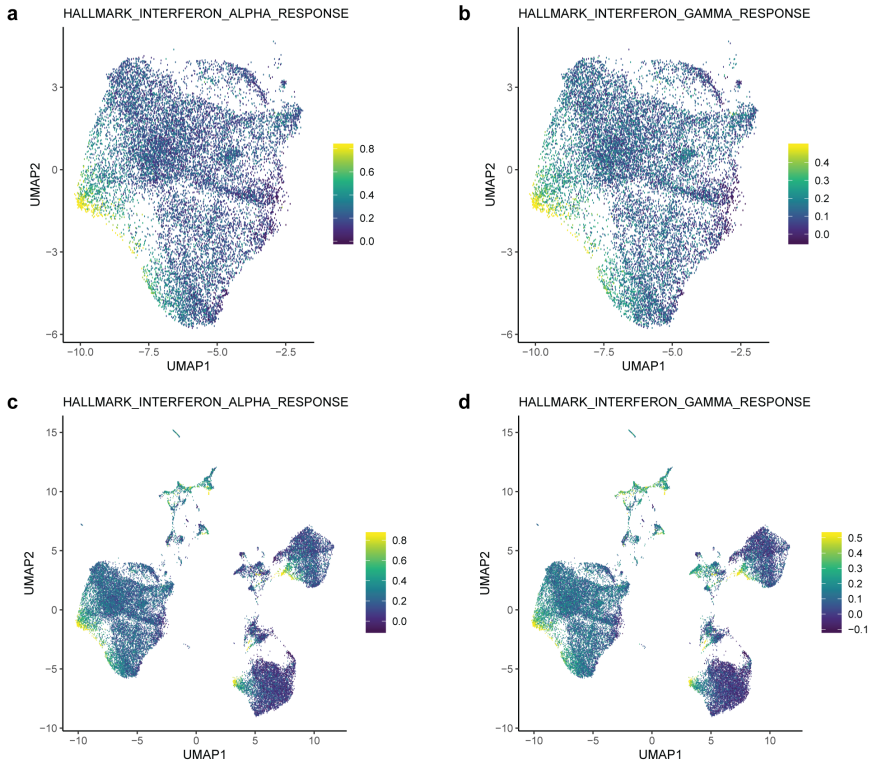
Supplementary Figure 2: UMAP embeddings of integration methods on spleen and lymph node data. a-e, For each method, UMAP plots colored by dataset, and by $\log(\text{counts} + 1)$ of key marker proteins (Supplementary Table 4).



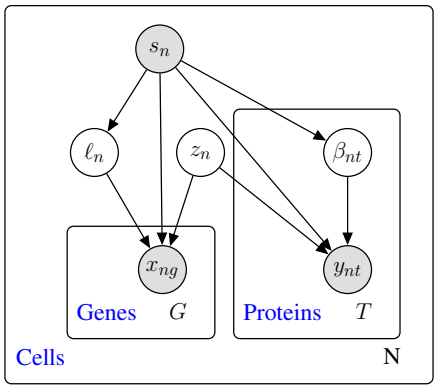
Supplementary Figure 3: Benchmarking of integration methods on PBMC data. Integration methods were applied to PBMC10k and PBMC5k. **a-e**, UMAP plots of integrated latent spaces. **f**, Latent mixing metric, feature retention metric, and clustering metric for each method (Methods). **g**, Measurement mixing metric applied individually to each batch-corrected feature (computed for $n = 4000$ genes and $n = 14$ proteins). Box plots indicate the median (center lines), interquartile range (hinges), whiskers at 1.5x interquartile range. **h**, UMAP plot of integrated latent space from totalVI union mode when holding out the proteins from PBMC5k. **i**, Root mean squared logarithmic error between imputed and observed proteins from PBMC5k for totalVI and Seurat v3. cTP-net did not provide predictions for CD127, CD15, CD25, PD-1, or TIGIT.



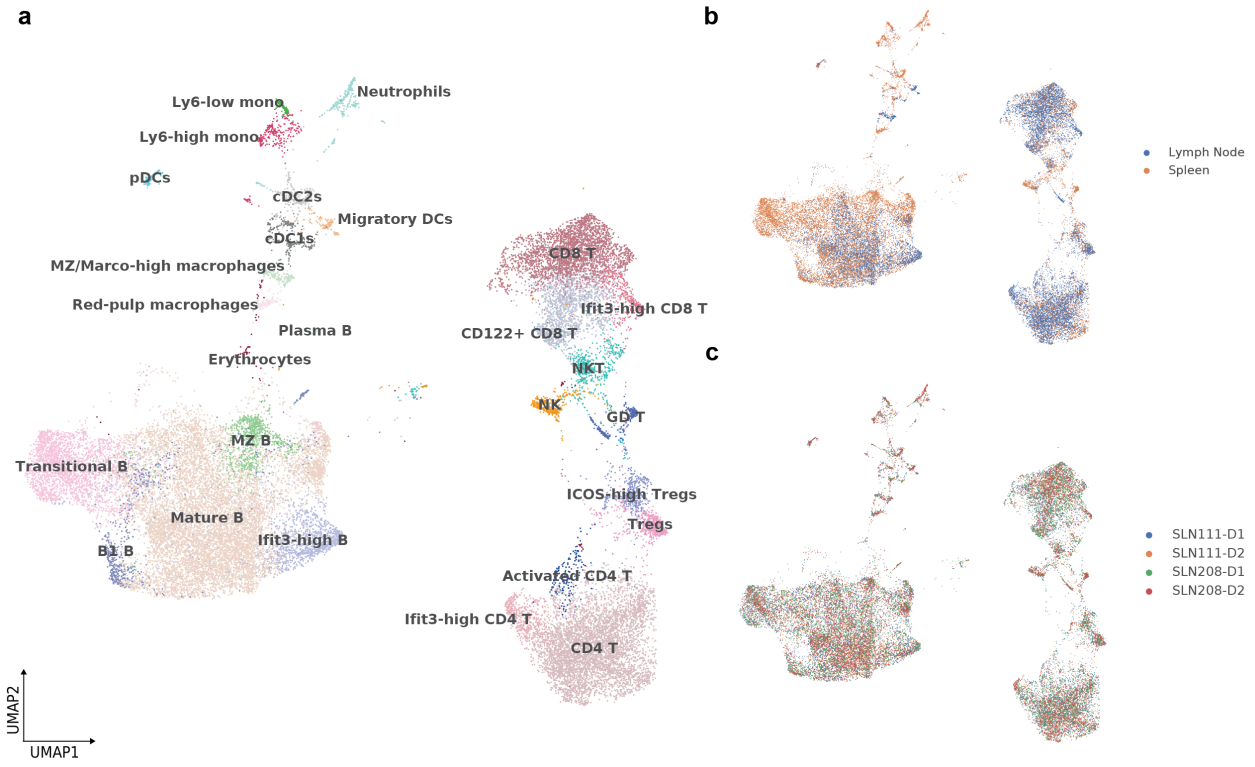
Supplementary Figure 4: Differential expression using totalVI-wBG, which does not correct for the protein background component. a Volcano plot for the ICOS-high Tregs vs CD4 T cells test. Putative positives and negatives are highlighted (Methods). **b** The LFC estimates for totalVI and totalVI-wBG on the CD4 T cells versus all others test.



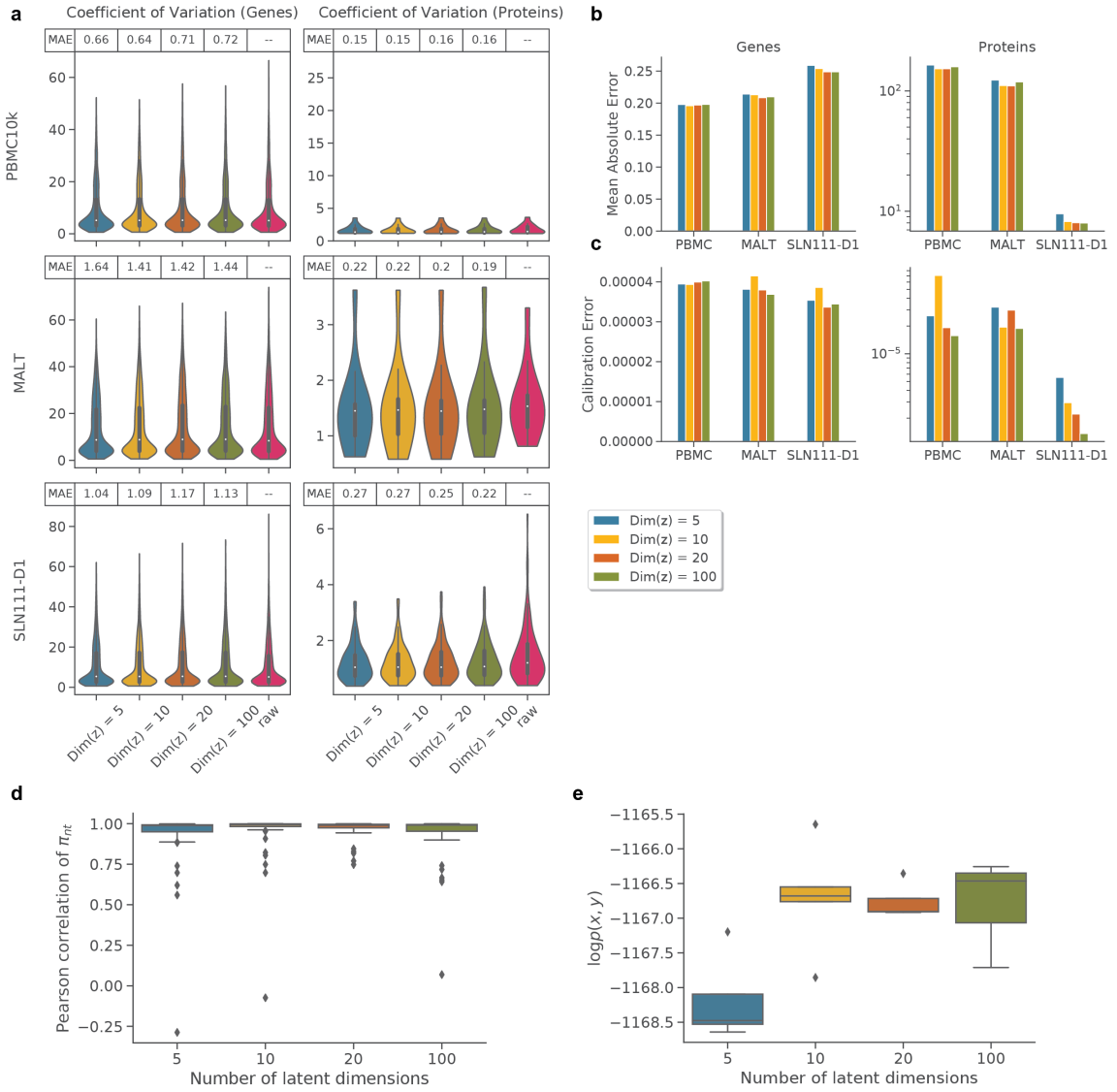
Supplementary Figure 5: Interferon signatures in the mouse spleen and lymph node. a, b, UMAP of totalVI latent space for B cells of the SLN-all dataset **(a)** colored by the Hallmark Interferon Alpha Response signature score and **(b)** colored by the Hallmark Interferon Gamma Response signature score (Methods). **c, d**, Same as in (a, b), but for all cells in SLN-all.



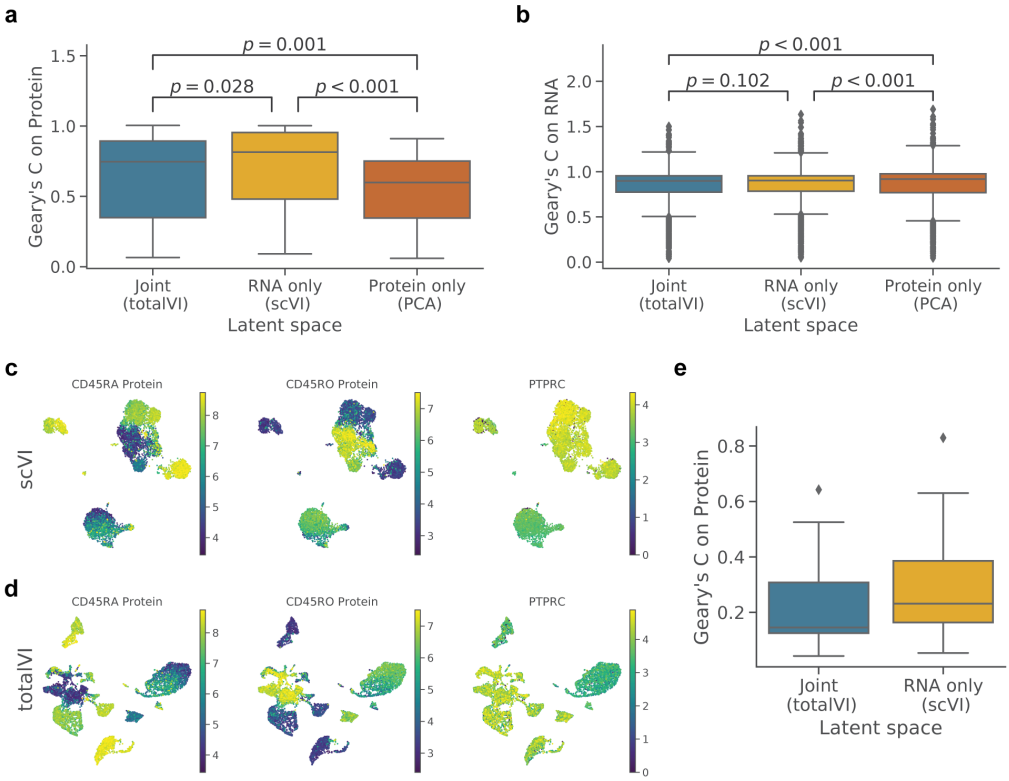
Supplementary Figure 6: totalVI probabilistic graphical model. Shaded nodes represent observed random variables. Unshaded nodes represent latent variables. Edges denote conditional independence. Rectangles ("plates") represent independent replication.



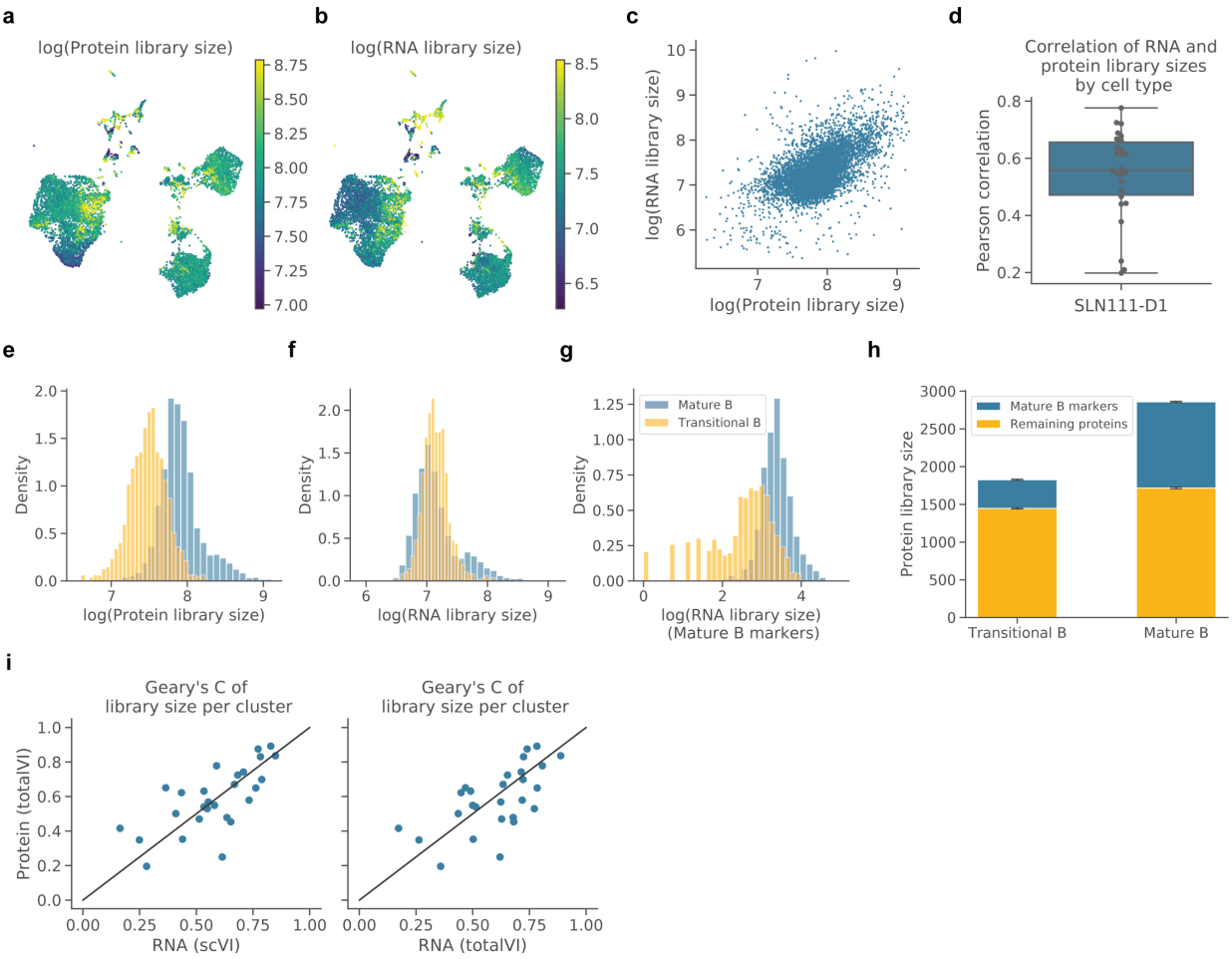
Supplementary Figure 7: Integration of SLN-all with union of panels. **a-c**, UMAP plot of SLN-all colored by **(a)** cell types derived from manual annotation of model run with intersection of panels, **(b)** tissue, and **(c)**, dataset.



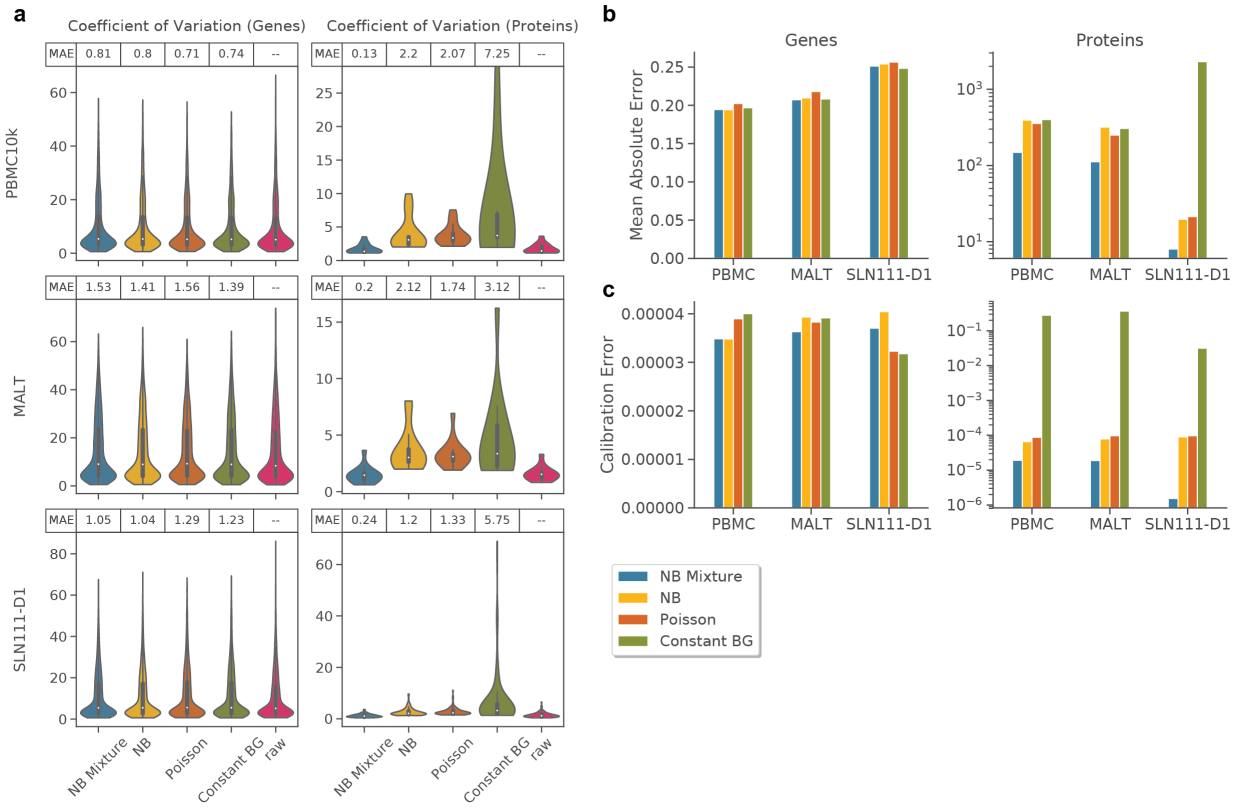
Supplementary Figure 8: Evaluation of totalVI across choice of number of latent dimensions. **a**, Posterior predictive check of coefficient of variation (CV) of genes and proteins. For totalVI with 5, 10, 20, and 100 latent dimensions, the average CV from posterior predictive samples was computed for each feature. Violin plots summarize the distribution of CVs for genes and proteins. Mean absolute error (MAE) between raw data CVs and average posterior predictive CV are reported. **b**, MAE between held out data and posterior predictive mean separated by genes and proteins for each model and dataset. **c**, Calibration error of held-out data reported separately for genes and proteins. **d**, Stability of estimate for background probability for each cell and protein with respect to default parameters on PBMC10k dataset ($n = 5$ model runs for each of the $n = 14$ proteins). **e**, Held-out marginal log likelihood on the PBMC10k dataset across latent dimensions ($n = 5$ model runs). Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5x interquartile range.



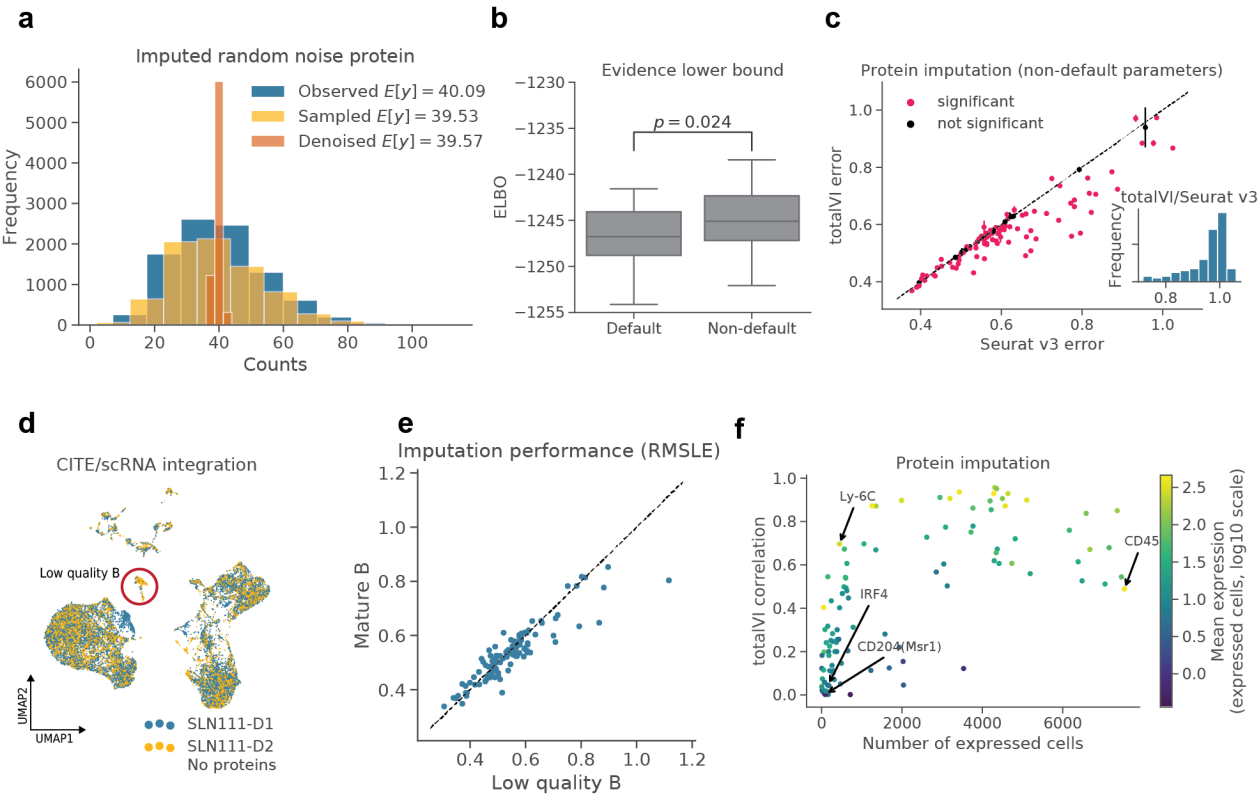
Supplementary Figure 9: Feature autocorrelation across latent space representations. a, b, Geary's C calculated in three latent spaces for each feature across all cells in the SLN111-D1 dataset. p-values indicate Wilcoxon rank sum test (two-sided). Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5x interquartile range. **(a)**, Geary's C for each protein ($\log(\text{protein counts} + 1)$, $n = 110$ proteins). In the RNA only vs protein only latent space comparison, $p = 4e - 6$. **(b)**, Geary's C for each gene (log library-size normalized; $n = 4005$ genes). In the RNA only vs protein only comparison, $p = 2e - 8$. For the totalVI joint latent space vs protein only comparison, $p = 1e - 12$. **c, d**, UMAP plots of CD45 isoform expression in the PBMC10k dataset in the scVI **(c)** or totalVI **(d)** latent space. CD45RA and CD45RO proteins are $\log(\text{protein counts} + 1)$ and *PTPRC* RNA is log library-size normalized. **e**, Geary's C calculated for each protein ($\log(\text{protein counts} + 1)$, $n = 14$ proteins) in the totalVI and scVI latent spaces depicted in (c, d). Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5x interquartile range.



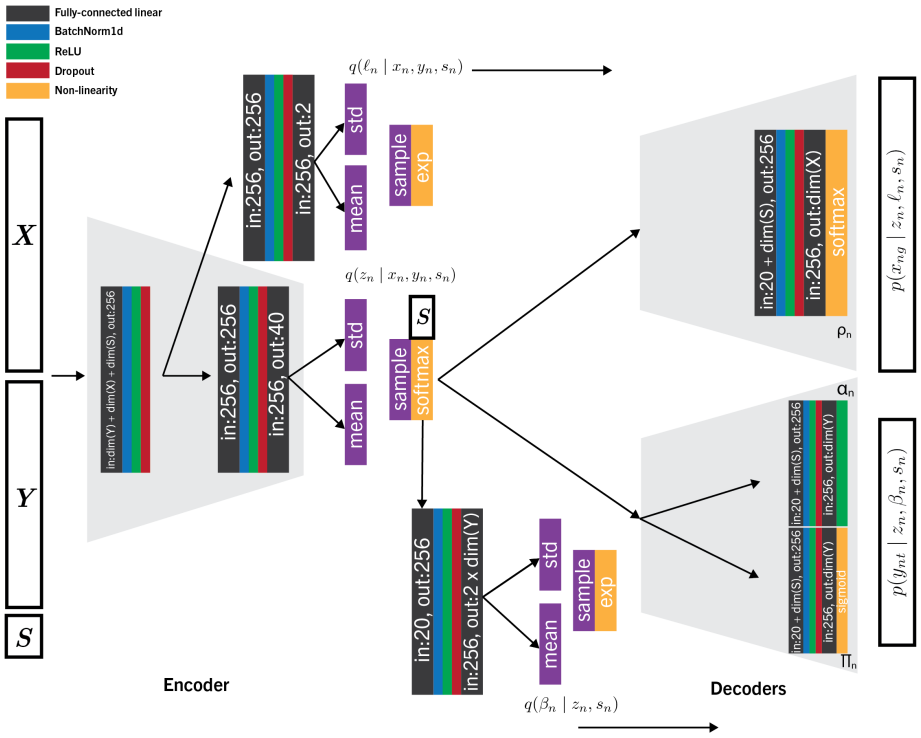
Supplementary Figure 10: Protein and RNA library sizes in the SLN111-D1 dataset. a, b, UMAP plots of SLN111-D1 cells in the SLN-all latent space colored by **(a)** log(protein library size) and **(b)** log(RNA library size). **c**, RNA library size vs protein library size (log scale) for each cell in the SLN111-D1 dataset. **d**, Pearson correlation of RNA and protein library sizes (log scale) by cell type ($n = 26$ cell types) for each cell type depicted in Fig. 4a excluding plasma B (two cells). Box plot indicates the median (center line), interquartile range (hinges), and whiskers at 1.5x interquartile range. **e-g**, Protein **(e)** and RNA **(f)** library sizes (log scale) for transitional and mature B cells in the SLN111-D1 dataset. **(g)** RNA library size (log scale) of the genes encoding the top 10 differentially expressed protein markers of mature B cells relative to transitional B cells. **h**, Protein library size in transitional and mature B cells separated by the top 10 differentially expressed markers of mature B cells and the remainder of proteins in the SLN111-D1 dataset. Data are presented as mean values +/- SEM ($n = 867$ transitional B cells and $n = 2,733$ mature B cells). **i**, Geary's C of library size computed per cluster in the SLN-all dataset compared across latent spaces. Left: RNA library size in the RNA-only scVI latent space vs protein library size in the totalVI latent space. Right: RNA library size in the totalVI latent space vs protein library size in the totalVI latent space.



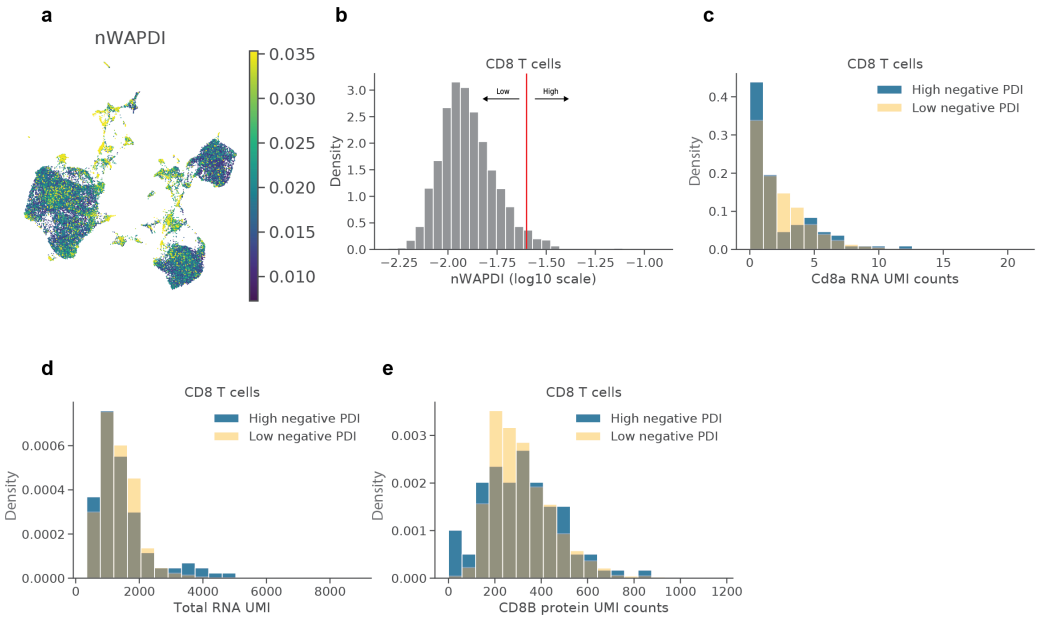
Supplementary Figure 11: Evaluation of totalVI across choice of protein likelihood. a, Posterior predictive check of coefficient of variation (CV) of genes and proteins. For totalVI with Poisson, negative binomial (NB), negative binomial mixture (NB Mixture), and negative binomial mixture with global background and library size correction (Constant BG) protein likelihood, the average coefficient of variation from posterior predictive samples was computed for each feature. Violin plots summarize the distribution of CVs for genes and proteins. Mean absolute error (MAE) between raw data CVs and average posterior predictive CV are reported. **b**, MAE between held out data and posterior predictive mean separated by genes and proteins for each model and dataset. **c**, Calibration error of held-out data reported separately for genes and proteins.



Supplementary Figure 12: Extended analysis of missing protein imputation. **a,** Distribution of observed protein counts (blue), totalVI imputed protein counts (denoised, orange), and samples from the totalVI imputed distribution (sampled, yellow). The random protein (observed, blue) is a simulated protein added to SLN111-D1 in the imputation task. **b,** Evidence lower bound for SLN imputation task using default parameters and updated (non-default) parameters across $n = 30$ trials. Significance was assessed with a two-sided Welch's t-test. Box plots indicate the median (center line), interquartile range (hinges), and whiskers at 1.5x interquartile range. **c,** totalVI imputation performance versus Seurat v3 imputation performance using non-default totalVI parameters. totalVI performance per protein is presented as mean RMSLE with error bars representing 95% confidence intervals of the mean estimate ($n = 30$ model initializations). Proteins colored in black are not significantly different in performance, while those in red are significantly different (two-sided Student's t-test, BH-adjusted p-value < 0.05). Inset displays ratio in performance between totalVI and Seurat v3. **d,** Reproduction of the UMAP in Figure 3f. **e,** Imputation performance (root mean squared log error) for each protein using only cells annotated as low quality B versus those annotated as mature B in SLN111-D2. **f,** totalVI imputation accuracy (Pearson correlation, log scale) versus number of expressed cells, which was estimated using the totalVI foreground probability ($\pi_{nt} < 0.5$). Points (proteins) are colored by mean expression in expressing cells.



Supplementary Figure 13: totalVI neural network architecture.



Supplementary Figure 14: Evaluation of SLN-all fit with negative widely applicable posterior dispersion indices (nWAPDI). a, nWAPDI computed for each cell in SLN-all on UMAP projection of the totalVI latent space. **b,** Distribution of nWAPDI for CD8 T cells, with decision boundary for cells marked as low nWAPDI and high nWAPDI. **c**, *Cd8a* expression (raw UMI counts) for CD8 T cells with low and high nWAPDI. **d**, RNA library size for same subpopulation split by low/high nWAPDI. **e,** Raw UMI counts of CD8B protein expression using same subpopulation and split.

Supplementary Note 1

On choosing the number of latent dimensions

The latent representation of a cell in totalVI, z_n, is critical to many of the tasks performed in this manuscript. Here we sought to understand the impact of the number of dimensions chosen for z_n on totalVI's modeling capabilities. To do so, we ran totalVI with 5, 10, 20, and 100 latent dimensions, and tested each model in our posterior predictive checking framework. For metrics computed on the training data and held-out data, totalVI's performance was fairly stable across this choice of hyperparameter (Supplementary Figure 8a-c). We note that the denoising component of totalVI also depends on z_n, so we measured the stability of the protein background probability estimate on the PBMC10k dataset across latent dimension choices and for five initializations each, using the default parameters (20 latent dimensions) as a baseline. We again found this estimate to be stable with respect to our default parameters (Supplementary Figure 8d). Finally, we measured the held out marginal log likelihood of the PBMC10k dataset across latent dimension choices. The marginal log likelihood improved for 100 dimensions, which could be due to increased capacity (Supplementary Figure 8e). We note that the relationship between the marginal log likelihood and all other downstream tasks is not well understood. While the marginal log likelihood on held-out data can be used a heuristic for choosing the number of dimensions, it remains unclear in this analysis what impact this will have on interpretation. This phenomenon of stability with respect to the number of latent dimension has also been reported by others [30–32].

Overall, we recommend using the default choice of 20 latent dimensions, which sits between the number used in the scVI [30] publication (10), and the number of principal components used in standard pipelines like Scanpy (about 30), and was used to achieve state-of-the-art results throughout this manuscript. Furthermore, the choice of 20 allowed us to interpret the totalVI latent space with archetypal analysis. If the number of dimensions is much smaller than the number of cell types, archetypal analysis may be more difficult.

Supplementary Note 2

On combining RNA and protein information in a joint latent space

To further explore the latent representation of totalVI's joint model and how it combines RNA and protein information to define cell-cell similarities, we considered the autocorrelation of each feature as measured by Geary's C [33]. We made comparisons to an RNA-only latent representation produced by scVI [30] and to a protein-only latent representation produced by PCA (each with 20 latent dimensions). In the SLN111-D1 dataset, we found that totalVI increased the autocorrelation (lower Geary's C) of protein features relative to RNA only ($p = 0.028$, Wilcoxon rank sum test (two-sided)) (Supplementary Figure 9a). Simultaneously, totalVI increased the autocorrelation of RNA features relative to protein only ($p< 0.001$) (Supplementary Figure 9b). Notably, the autocorrelation of RNA features in the totalVI latent space was not significantly different from that of scVI ($p = 0.102$), implying that the joint latent space of totalVI does not sacrifice the richness of information in the RNA data in order to incorporate protein information.

Combining RNA and protein information into a joint latent representation can be particularly beneficial when the protein data includes isoforms not detectable by RNA sequencing that measures transcript counts rather than full length molecules. For example, the CD45 isoforms CD45RA and CD45RO are commonly used surface markers to distinguish naive from memory human T cells. In an RNA-based analysis, it is possible to define a latent space based on RNA followed by the annotation of cells with their protein isoform expression (Supplementary Figure 9c). However, a joint analysis can make use of protein isoform expression along with transcriptome measurements to define a latent space and cell-cell similarities (Supplementary Figure 9d). As expected, totalVI's joint model increases the autocorrelation of proteins relative to an RNA-only analysis (Supplementary Figure 9e), indicating that totalVI incorporates isoform information into the latent space.

Supplementary Note 3

Protein considerations

Experimental considerations Sources of technical variation in CITE-seq experiments, particularly protein background, are dependent on the experimental method itself. There are a number of potential experimental sources of background. We primarily discussed ambient antibodies and non-specific antibody binding. Another potential source of background could arise from oligonucleotide barcodes that become dissociated from their conjugated antibody. Similarly to barcoded antibodies, ambient oligonucleotide barcodes could contaminate cell-containing droplets or could non-specifically bind to the cell surface. In this study, we do not distinguish between background due to antibodies or background due to free oligonucleotide barcodes.

Although in our experiments we used the standard CITE-seq protocol, there are a number of protocol modifications that could change the amount of background. For instance, increasing the number of washes after staining cells with antibodies could reduce ambient background. Alternatively, a buffer modification could reduce the amount of non-specific binding. Both washing and blocking are frequently considered in flow cytometry protocol designs. However, implementing these protocol changes in an effort to eliminate background could come with trade-offs; reducing background by washing and blocking would likely reduce true signal by reducing total cell numbers or blocking specific binding, respectively.

Another common experimental practice to modulate the amount of background is antibody titration, meaning that different antibodies are added to the experiment at different concentrations. At the optimal titration, an antibody would have the maximal signal-to-noise ratio. This would require the antibody to be present at a sufficient concentration to specifically bind its target protein and generate a detectable signal, but not at so high a concentration as to increase protein background by binding non-specifically or remaining at high concentration in the ambient solution. In a CITE-seq experiment, it is possible that the recommended antibody concentration is too low to detect foreground signal from a given protein. Finding the optimal concentration for each antibody might be challenging, since the optimal concentration might be different for different cell types or experimental systems. If titrations are modified per antibody, there are a few points to consider. When antibodies are titrated at different concentrations, it becomes infeasible to quantify absolute protein levels. For example, it would not be possible to determine if one protein was expressed at a higher level than another protein. In addition, even if every protein in the cell were measured with a theoretical unbiased antibody panel, the sum of all protein counts from a cell (referred to as the protein library size) could not serve as a meaningful estimate of total protein molecules in a cell or cell size because the relative amounts of each protein have been manipulated.

For proteins that are expressed at low levels or are only expressed in rare cell types, it might be necessary to increase sequencing depth to increase the sensitivity to detect these molecules. In addition, sequencing depth could play a role in the ability of totalVI to decouple protein foreground and background (i.e., with more counts, it might become easier to separate what appear to be overlapping foreground and background distributions). Determining the optimal sequencing depth for protein panels could be an important cost consideration in CITE-seq experiments, particularly as the size of protein panels increases. Since in the CITE-seq protocol RNA and protein libraries are prepared independently, future work could determine the value of these two molecules in various downstream analysis tasks to make recommendations for sequencing depth for each library.

Because the barcoded antibodies used in this study came from clones that have been previously validated, we were surprised to find that some common protein markers (e.g., IgM, CCR7) appeared to have little or no signal. Aside from the consideration of titration and sequencing depth discussed above, an additional explanation could be the uniform staining conditions for all antibodies in the CITE-seq panel simultaneously. For example, the chemokine receptor CCR7 is a well-documented marker in T cells and typically requires staining at higher temperature and for longer times than other antibodies due to its constant cycling onto and off of the cell surface. For future CITE-seq experiments, it might be worthwhile to consider the optimal staining conditions (e.g., time and temperature) for each antibody independently rather than staining with all antibodies at once.

Modeling considerations Guided by the points raised above, we considered a variety of protein likelihoods before settling on the version used in this manuscript. Among our considerations were the interpretability of the parameters as well as how well the likelihood captured our view on the CITE-seq protein data generating process.

In our modeling and analysis, we considered whether protein library size should be taken into account. For example, we considered models that included a latent variable for the protein library size (analogous to ℓ_n for RNA). Here, we discuss why protein library size does not convey the same information as RNA library size and is thus not treated the same in the totalVI model.

In scRNA-seq data, we consider library size to be a nuisance factor that is reflective of a combination of sequencing depth and cell size. Although the number of RNA transcripts and the number protein molecules both scale with the size of a cell [34], the unbiased sampling of the transcriptome is much more likely to approximate the relative size of a cell than the sampling of a limited selection of proteins on the cell surface. In CITE-seq experiments where only selected markers are measured, there is no guarantee that the markers selected are representative of the total protein content in the cell. For example, a cell with few detected protein counts might in reality express other unmeasured proteins at higher levels, meaning this cell's total counts reflect the selection of markers rather than reflecting nuisance variation like sequencing depth or cell size. Therefore, treating protein library size as a nuisance factor does not make sense in this context. Because measured proteins are biased in this manner, we do not assume that the protein data is compositional, as is assumed by other methods that use a centered log ratio (CLR) transformation for normalization per cell [35].

To further explore the effects of library size, we considered protein and RNA library sizes in the SLN111-D1 dataset (Supplementary Fig. 10a-c). While RNA and protein library sizes (log scale) were positively correlated (Pearson correlation of 0.54), this correlation varied widely by cell type (Supplementary Fig. 10d). This observation suggested that within some cell types, protein library size might approximate a cell's relative size similarly to RNA library size. However, in other cell types, bias in the antibody panel results in total protein expression that reflects biological differences in the measured proteins rather than a global measurement of protein content. The difference between RNA and protein library size can be observed when comparing transitional and mature B cells, which have significantly different protein library sizes (Welch's t-test, $p < 0.01$), but not significantly different RNA library sizes (Welch's t-test, $p = 0.098$) (Supplementary Fig. 10e-f). The difference in protein library size between these cell types can largely be explained by differences in expression of measured mature B cell markers: 74% of the mean difference in protein library size is driven by the 10 most differentially expressed proteins in mature vs transitional B cells (Methods), including known markers like IgD and CD23 (Supplementary Fig. 10g, h). Therefore, normalizing each cell by its protein library size would remove biological differences and introduce additional bias based on the selection of measured proteins. Finally, we considered how library size is reflected in the latent space by calculating the autocorrelation of library size per cluster as measured by Geary's C [33]. The autocorrelation of protein library size in the totalVI latent space was similar to the autocorrelation of RNA library size in either the RNA-only latent space computed by scVI or in the totalVI latent space, both of which remove the effect of RNA library size (Supplementary Fig. 10i). This finding further supported the notion that the totalVI latent space was not unduly biased by protein library size. In the future, more unbiased protein panels might necessitate further consideration of protein library size in CITE-seq data analysis.

We also considered whether RNA background should be addressed similarly to protein background in our model. In addition to our observations that levels of background RNA were far lower than protein background (Extended Data Fig. 3d-f), we considered estimates of UMI counts due to background for RNA and protein. For RNA background, we refer to the DecontX method that estimates and removes contamination of ambient RNA in scRNA-seq data [36]. According to the DecontX study, in three scRNA-seq datasets of different tissues collected using the 10x v3 platform, the median percentage of RNA UMI counts estimated to be derived from ambient RNA was 0.03% (brain), 0.12% (heart), and 0.56% (PBMC) [36]. For comparison, we estimated the number of protein UMI counts that could be attributed to background in each cell by summing all protein UMI counts with a foreground probability < 0.5 and dividing by the total protein UMI counts. For the data reported in this study (also collected using the 10x v3 platform), the estimated median percentage of protein background UMI counts across all cells in SLN-all was 12.71%. By these estimates, RNA background UMI counts are low (<1%, approximately two orders of magnitude lower than protein background UMI counts). We therefore chose not to explicitly model RNA background in totalVI.

Additionally, we considered alternative models to decouple protein background. Initially, we attempted to use simple likelihoods like Poisson or negative binomial, or models that assume every cell receives the same distribution of protein background scaled by some cell-specific scalar. However, we found these models inadequate for decoupling the protein signal, which again suggests that ambient antibodies can not fully explain the protein background, and that our proposed negative binomial mixture fits the data better (Supplementary Fig. 11).

The likelihood we used in this manuscript,

$$y_{nt} \mid z_n, \beta_n, s_n \sim \text{NegativeBinomialMixture}(\beta_{nt}, \beta_{nt}\alpha_{nt}, \pi_{nt}), \tag{1}$$

also has some important downstream considerations. First, the mixture assumes that the observed counts for a given cell n and protein t are generated from either the background component (with probability π_{nt}) or the foreground component (with probability $1 - \pi_{nt}$). Despite the fact that the background mean parameter β_{nt} appears in the foreground mean $\beta_{nt}\alpha_{nt}$, this likelihood does not allow us to correct the foreground for possible background contamination. Here, the double usage of β_{nt} is to help identify the mixture model. In other words, we cannot "subtract the background" from y_{nt} that are determined to be in the foreground. Perhaps this limitation could be addressed in future work, in which different latent variables are associated with local and global sources of background; though this will require greater understanding of the experimental mechanisms previously discussed.

As a final point, the negative binomial mixture we used in this manuscript may be less suitable in the case where the dataset contains a homogeneous population of cells. This is due to there being a lack of cells that would be considered background for the proteins. In this case, one can either use the mean of the negative binomial mixture (without expected background subtraction) for downstream analysis, or optionally use the negative binomial distribution as the protein likelihood (an option in totalVI).

Supplementary Note 4

On imputing missing proteins

In our analysis, we have shown that totalVI can accurately predict missing proteins, with improvement over the predictions of Seurat v3. Here we further discuss the merits and limitations of missing protein imputation.

We first sought to quantify any bias in totalVI's protein imputation. As an illustrative example, we simulated a protein as independently and identically distributed from a negative binomial distribution (`np.random.negative_binomial(10, 0.2)`.). We concatenated this protein to the others and reran the imputation example used in Fig. 3 on our spleen and lymph node data (SLN111-D1 and SLN111-D2 with no proteins). The default totalVI prediction is displayed as the orange histogram in Figure 12a. As the totalVI denoised counts represent the expected value, we should expect that if totalVI does not produce biased predictions that the values pileup around 40, the expected value of the simulated negative binomial data (in blue). Because totalVI fits the full distribution, instead of reporting the expectation, we can sample from it. From this sampling, we observed that the samples closely matched the empirical distribution of the random protein. Taken together, in this particular experiment, totalVI's predictions have little bias, but the degree of noise in the proteins sets a ceiling on the imputation performance of any algorithm.

We also sought to understand the impact of totalVI's hyperparmeters on the imputation task. Throughout the manuscript we used a set of default hyperparameters, so as the provide reasonable defaults that would yield good performance across many tasks and datasets with distinct characteristics. However, the task of protein imputation requires the model to not backpropagate certain nodes in the network as well as to handle zeros for missing data; so, it is reasonable that another set of hyperparameters could yield improved performance on this task. Guided by maximizing the evidence lower bound, we found another set of hyperparameters that differed from the defaults only in the number of hidden layers (two, relative the default one), and a reduction of the learning rate from `4e-3` to `2e-3`. We again performed the imputation task on the spleen and lymph node data over 30 initializations. We found the evidence lower bound of the data to be significantly higher with the non-default parameters (Welch's t-test, p-value < 0.05; Supplementary Figure 12b). Furthermore, we found 89 proteins to be significantly different in their root mean squared logarithmic error to that of Seurat v3, 68 of which had better performance for totalVI (Supplementary Figure 12c). This is an improvement over the default parameters, though it is quite minor, indicating that the task is somewhat robust to these choices.

Finally, we consider the usage of totalVI imputed proteins as proxies for real observed protein measurements. In our analysis, totalVI had good imputation accuracy for many proteins. One relevant consideration is the quality of the proteins in the reference dataset. Imputation performance may suffer due to poor antibodies or lack of biological expression, and it will not always be straightforward to understand which effect is being observed, especially as CITE-seq panels grow to be more unbiased. Another consideration is the extent to which the reference and query datasets share underlying biology. totalVI relies on the assumption that a cell with no observed proteins will have cells from the dataset with observed protein expression in its neighborhood in the latent space. If a cell type is missing from the reference dataset that is observed in the query dataset, we do not expect good imputation performance; though it may depend how biologically similar the missing cell type is to the others in the reference dataset.

For example, the UMAP plot in Fig. 3f (reproduced here as Supplementary Figure 12d) revealed a small subpopulation of cells from SLN111-D2 that did not integrate well with the cells of SLN111-D1. Based on our annotations, these cells were overwhelmingly from a subpopulation of low quality B cells (high mitochondrial content) that were enriched in the SLN111-D2 dataset relative to the remaining spleen and lymph node datasets, explaining the lack of mixing. We compared the imputation performance via correlation for these cells relative to a related population of mature B cells, and found the performance tended to be similar in the low quality B cells (Supplementary Figure 12e). However, this result may be due to the relative similarity of these low quality B cells to the higher quality mature B cells, coupled with the fact that the RNA and not the protein data were low quality. Generally, we believe that imputation performance will be better for those cells that mix well with CITE-seq cells. The potential pitfall of poor imputation quality can be mitigated by empirically quantifying the presence of CITE-seq cells in a cell's neighborhood. While imputed proteins can be used in downstream analysis to generate hypotheses, these predictions should be validated experimentally before drawing concrete biological conclusions.

We also note that totalVI is capable of imputing the expression of proteins that are expressed in rare cell types (Supplementary Figure 12f). Here we investigated the Pearson correlation as an imputation performance

metric, which allowed us to evaluate proteins on the same scale, as the RMSLE depends on the range of expression values and mean of the protein. In Supplementary Figure 12f, we can see that totalVI imputes Ly-6C well, which is expressed in smaller subpopulations of CD8 T cells and monocytes. In contrast, IRF4 is a negative control protein (intracellular) that was detected in very few cells (likely technical artifacts), explaining its low correlation.

Supplementary Note 5

Integrating out latent variables

Here we show that if

$$w \sim \text{Gamma}(\theta, \ell\rho) \tag{2}$$
$$x \mid w \sim \text{Poisson}(w) \tag{3}$$

then $x \sim \text{NegativeBinomial}(\ell\rho, \theta)$. Note that we have dropped all subscripts and each variable here is a scalar. While we parameterize the Gamma with its shape and mean, a more conventional form is with its shape and rate, so $w \sim \text{Gamma}(\theta, \theta/(\ell\rho))$

$$p(x) = \int_0^\infty p(x \mid w)p(w)dw \tag{4}$$
$$= \int_0^\infty \frac{w^x e^{-w}}{\Gamma(x+1)} \frac{(\theta/(\ell\rho))^\theta}{\Gamma(\theta)} w^{\theta-1} e^{-\theta w/(\ell\rho)} dw \tag{5}$$
$$= \frac{(\theta/(\ell\rho))^\theta}{\Gamma(x+1)\Gamma(\theta)} \int_0^\infty w^{x+\theta-1} e^{-(1+\theta/(\ell\rho))w} dw \tag{6}$$
$$= \frac{(\theta/(\ell\rho))^\theta}{\Gamma(x+1)\Gamma(\theta)} \frac{\Gamma(x+\theta)}{(1+\theta/(\ell\rho))^{x+\theta}} \tag{7}$$
$$= \frac{\Gamma(x+\theta)}{\Gamma(x+1)\Gamma(\theta)} \left(\frac{\theta/(\ell\rho)}{1+\theta/(\ell\rho)}\right)^\theta \left(\frac{1}{1+\theta/(\ell\rho)}\right)^x \tag{8}$$
$$= \frac{\Gamma(x+\theta)}{\Gamma(x+1)\Gamma(\theta)} \left(\frac{\theta}{\ell\rho+\theta}\right)^\theta \left(\frac{\ell\rho}{\ell\rho+\theta}\right)^x \tag{9}$$

In the fourth line, we use the fact that the integrand is an unnormalized gamma distribution. The final line is a negative binomial distribution with mean $\ell\rho$ and inverse dispersion θ. Therefore, we have a direct link between the parameters of the negative binomial and the underlying parameters of the Poisson rate. Finally, we note that we could have written $w \sim \text{Gamma}(\theta, \rho)$ and $x \mid w \sim \text{Poisson}(\ell w)$ and achieved the same result.

Supplementary Note 6

totalVI implementation details

Evidence lower bound derivation Here we derive the Evidence Lower Bound (ELBO), which is ultimately used in optimizing the model and variational parameters. For shorthand, we drop subscripts and inference and generative parameters ν and η. The joint likelihood based on the totalVI generative model for a single cell factorizes as

$$p(x, y, \beta, z, \ell \mid s) = p(x \mid z, \ell, s)p(y \mid \beta, z, s)p(\beta \mid s)p(z)p(\ell \mid s). \tag{10}$$

In the model specification, we use the latent variable $z \sim \text{LogisticNormal}(0, I)$. Here we use the logistic normal definition of [37, 38], in which a normal random variable $\delta \sim \text{Normal}(0, I)$ is transformed by a softmax function, embedding the random variable in the simplex. Thus, $z = \text{softmax}(\delta)$. However, the softmax function is not invertible, so for simplicity we consider the underlying latent variable δ. In this setting, z, which is ultimately the input to the decoder, is treated as a likelihood parameter. Therefore, we can rewrite the joint likelihood as

$$p(x, y, \beta, \delta, \ell \mid s) = p(x \mid \delta, \ell, s)p(y \mid \beta, \delta, s)p(\beta \mid s)p(\delta)p(\ell \mid s). \tag{11}$$

To perform variational inference, we define the variational posterior distribution as

$$q(\beta, \delta, \ell \mid x, y, s) = q(\beta \mid \delta, s)q(\delta \mid x, y, s)q(\ell \mid x, y, s). \tag{12}$$

The ELBO is derived using Jensen's inequality. We use the shorthand notation $q(\beta, \delta, \ell) = q(\beta, \delta, \ell \mid x, y, s)$.

$$\begin{aligned}
\log p(x, y \mid s) =& \log \mathbb{E}_{q(\beta,\delta,\ell)} \left[\frac{p(x, y, \beta, \delta, \ell \mid s)}{q(\beta, \delta, \ell)} \right] && (13)\\
\geq& \mathbb{E}_{q(\beta,\delta,\ell)} \left[\log \frac{p(x, y, \beta, \delta, \ell \mid s)}{q(\beta, \delta, \ell)} \right] && (14)\\
=& \mathbb{E}_{q(\beta,\delta,\ell)} \left[\log p(x, y \mid \beta, \delta, \ell, s) \right] + \mathbb{E}_{q(\beta,\delta,\ell)} \left[\log \frac{p(\beta \mid s)p(\delta)p(\ell \mid s)}{q(\beta, \delta, \ell)} \right] && (15)\\
=& \mathbb{E}_{q(\beta,\delta,\ell)} \left[\log p(x, y \mid \beta, \delta, \ell, s) \right] - \text{KL}\left(q(\ell) \parallel p(\ell \mid s) \right) - \text{KL}\left(q(\delta) \parallel p(\delta) \right) && (16)\\
& - \mathbb{E}_{q(\delta)} \left[\text{KL}\left(q(\beta) \parallel p(\beta \mid s) \right) \right] && (17)
\end{aligned}$$

To compute the KL divergences of lognormal random variables we note that the KL divergence is invariant to invertible transformations, so the KL can be computed in closed form using the KL between normal random variables. The log likelihood of a negative binomial mixture distribution is computed using numerically stable functions in Pytorch (see below). The ELBO derived here is amenable to the reparameterization trick used to train VAEs [39]. Estimates of the expectations in the ELBO are taken via Monte Carlo and noisy gradients of the ELBO are used in a stochastic optimization scheme. A sketch of the inference procedure for totalVI is in Algorithm 2.

Approximate posterior specification The approximate posterior distributions are specified by neural networks. In particular, one neural network takes as input the triple (x, y, s) and outputs the parameters of $q(\delta \mid x, y, s)q(\ell \mid x, y, s)$. An additional neural network maps (δ, s) to the mean and variance parameters of $q(\beta \mid z, s)$ through z. The variational distributions match their priors in family (e.g., $q(\delta \mid x, y, s)$ is a Gaussian with diagonal covariance matrix). A posterior draw of z, which we used as input to clustering and visualization algorithms, as well as used for archetypal analysis is then obtained by

1. Draw δ from $q(\delta \mid x, y, s)$
2. Set $z = \text{softmax}(\delta)$

The mean of $q(z \mid x, y, s)$ can be computed using Monte Carlo integration.

Neural networks The encoder neural network has one shared hidden layer of 256 nodes followed by a layer of 512 nodes. The output of the 512 nodes are split in half and are used as input for linear layers that parameterize $q(\delta \mid x, y, s)$ and $q(\ell \mid x, y, s)$, respectively. The final encoder neural network of one hidden layer and 256 hidden nodes takes as input (z, s) and outputs the parameters of $q(\beta \mid \delta, s)$. The parameters of δ are 20-dimensional mean and variance parameters. The decoder consists of three individual neural

Algorithm 2: Inference for totalVI

Initialize inverse dispersion parameters θ, ϕ, background parameters c_t, d_t and encoder/decoder neural network parameters.
for *iteration* $i = 1, 2, ...,$ **do**
  Randomly choose M cells for mini-batch $\mathcal{C}$
  for *each cell* n *in* $\mathcal{C}$ **do**
    Encode x_n, y_n, s_n to obtain approximate posterior parameters
    Sample z_n, ℓ_n, β_n from approximate posterior $q(\beta_n, \delta_n, \ell_n \mid x_n, y_n, s_n)$
    Decode z_n, s_n to obtain likelihood parameters α_n, π_n, ρ_n
    for *each gene* g **do**
      Compute $\log p(x_{ng} \mid \ell_n, z_n, s_n)$
    for *each protein* t **do**
      Compute $\log p(y_{nt} \mid \beta_{nt}, z_n, s_n)$
  Update parameters using gradient of ELBO estimate

networks each with one hidden layer and 256 nodes. The first maps to the parameters of the mean of the RNA likelihood (ρ_n). The second maps to the foreground mean of the protein likelihood (α_n). The third maps to the mixing parameter of the protein likelihood mixture (π_n). Each of these decoder networks takes as input (z, s). Furthermore, (z, s) are reinjected at each hidden layer. All neural networks use batch normalization [40], dropout regularization [41], and ReLU activations in hidden layers. The model parameters θ, ϕ, c, and d are treated as global neural network parameters, optimized to maximize the ELBO. A schematic of the architecture is in Supplementary Figure 13. We note that the architecture (fully-connected, one to two hidden layers, ReLU activation, etc.) is similar to other VAE/autoencoder models used for single-cell data including the scVI and DCA models [30, 31].

Hyperparameters The neural network architecture previously described was used throughout this manuscript without modification. There are a number of other hyperparameters used to train neural networks that we also held constant in all experiments. This includes the learning rate of the optimizer (`lr=4e-3`), the size of the training test (90%), the KL warmup scheme ($0.75\times$`NumCells` minibatches, or 213 epochs with the fixed training set size of 90%), and the number of training epochs (500 epochs). An early stopping scheme is performed with respect to the 10% of cells in the test set. If there is no improvement of the held-out ELBO of the test set with 30 epochs, the learning rate is multiplied by 0.6. If there is no improvement after 45 epochs, the inference procedure is stopped.

Numerical considerations for the negative binomial mixture distribution For a single negative binomial mixture component, we use numerically stable functions provided by PyTorch (e.g., the log gamma function). For a mixture of negative binomials, we rewrite the distribution to use numerically stable functions like `logsumexp` and `softplus`.

Let $p^b(y) = p(y \mid z, \beta, s, v = 1)$ be the probability mass function for the background and $p^f(y) = p(y \mid z, \beta, s, v = 0)$ be the probability mass function for the foreground. Then by integrating over v (recalling that $v \sim \text{Bernoulli}(\pi)$),

$$\log p(y \mid z, \beta, s) = \log \big(\pi p(y \mid z, \beta, s, v = 1) + (1 - \pi) p(y \mid z, \beta, s, v = 0)\big). \tag{18}$$

We now rewrite this in a form more amenable for optimization. We recall from Algorithm 1 that $\pi = h_\pi(z, s; \Omega)$. Thus, with $c(z, s) = \text{logit}(h_\pi(z, s))$, then $\pi = 1/(1 + \exp(-c(z, s)))$. Also, let $\mathcal{S}$ be the softplus function: $x \to \log(1 + e^x)$.

$$\begin{aligned}
\log p(y \mid z, \beta, s) &= \log \big(\pi p^b(y) + (1 - \pi) p^f(y)\big) && (19)\\
&= \log \big(p^f(y) + \pi p^b(y) - \pi p^f(y)\big) && (20)\\
&= \log \left(\frac{p^f(y) + p^f(y) e^{-c(z,s)}}{1 + e^{-c(z,s)}} + \frac{p^b(y) - p^f(y)}{1 + e^{-c(z,s)}}\right) && (21)\\
&= \log \left(p^b(y) + p^f(y) e^{-c(z,s)}\right) - \log \left(1 + e^{-c(z,s)}\right) && (22)\\
&= \log \left(e^{\log p^b(y)} + e^{\log p^f(y)} e^{-c(z,s)}\right) - \log \left(1 + e^{-c(z,s)}\right) && (23)\\
&= \text{logsumexp} \left(\log p^b(y), \log p^f(y) - c(z, s)\right) - \mathcal{S}(-c(z, s)) && (24)
\end{aligned}$$

Supplementary Note 7

Posterior dispersion indices highlight model misfit

Here we describe how we used posterior dispersion indices to further evaluate the totalVI model fit. In this analysis, we used the negative widely applicable dispersion index (nWAPDI) [42]. With this metric, each cell gets a value describing the uncertainty of the likelihood of a particular cell with respect to the latent variables. Higher values indicate that the model is failing to explain the observed expression in a particular cell. The nWAPDI is computed for cell n as

$$\text{nWAPDI(n)} := -\frac{\sigma^2_{\log}(n)}{\log \mu(n)}, \tag{25}$$

where the quantities

$$\mu(n) = \mathbb{E}_{q(z_n,\beta_n,\ell_n)}[p(x_n, y_n \mid z_n, \beta_n, \ell_n, s_n)], \tag{26}$$

$$\sigma^2_{\log}(n) = \mathbb{V}_{q(z_n,\beta_n,\ell_n)}[\log p(x_n, y_n \mid z_n, \beta_n, \ell_n, s_n)], \tag{27}$$

can be computed using samples from the approximate posterior (1000 samples for each cell). We computed the nWAPDI for each cell in the SLN-all model fit (Supplementary Figure 14a). Next, we looked at the distribution of nWAPDI scores inside a homogeneous cell type like CD8 T cells and explored phenotypic differences between outliers and the remaining CD8 T cells. We found that those cells with high nWAPDI tended to have a surprising zero UMI count frequency for key cell type markers like the Cd8a gene and the CD8B protein, and it could not be explained by a difference in library size (Supplementary Figure 14b-e). This indicates that totalVI may not be explaining particularly surprising zeros well and could suggest the use of zero-inflated distributions, though this remains future work.

References

1. Li, B. *et al.* Cumulus provides cloud-based data analysis for large-scale single-cell and single-nucleus RNA-seq. *Nature Methods* **17,** 793–798 (2020).
2. Kirchner, J. & Bevan, M. J. ITM2A is induced during thymocyte selection and T cell activation and causes downregulation of CD8 when overexpressed in CD4+CD8+ double positive thymocytes. *Journal of Experimental Medicine* **190,** 217–228 (1999).
3. Kreslavsky, T. *et al.* Essential role for the transcription factor Bhlhe41 in regulating the development, self-renewal and BCR repertoire of B-1a cells. *Nature Immunology* **18,** 442–455 (2017).
4. Macias-Garcia, A. *et al.* Ikaros and B1 cells Ikaros is a negative regulator of B1 cell development and function. *Journal of Biological Chemistry* (2016).
5. Shi, Z. *et al.* Human CD8+ CXCR3+ T cells have the same function as murine CD8+ CD122+ Treg. *European Journal of Immunology* **39,** 2106–2119 (2009).
6. Liu, J., Chen, D., Nie, G. D. & Dai, Z. CD8+CD122+ T-cells: A newly emerging regulator with central memory cell phenotypes. *Frontiers in Immunology* **6** (2015).
7. Lai, L., Alaverdi, N., Maltais, L. & Morse, H. C. Immunophenotyping mouse cell surface antigens: Nomenclature and immunophenotyping. *The Journal of Immunology* (1998).
8. Merad, M., Sathe, P., Helft, J., Miller, J. & Mortha, A. The Dendritic cell lineage: ontogeny and function of dendritic cells and their subsets in the steady state and the inflamed setting. *Annual Review of Immunology* **31,** 563–604 (2013).
9. Durai, V. & Murphy, K. M. Functions of murine dendritic cells. *Immunity* **45,** 719–736 (2016).
10. Scott, R. E., Ghule, P. N., Stein, J. L. & Stein, G. S. Cell cycle gene expression networks discovered using systems biology: Significance in carcinogenesis. *Journal of Cellular Physiology* **230,** 2533–2542 (2015).
11. Doty, R. T. *et al.* Coordinate expression of heme and globin is essential for effective erythropoiesis. *Journal of Clinical Investigation* **125,** 4681–4691 (2015).
12. Sagar *et al.* Deciphering the regulatory landscape of $\gamma\delta$ T Cell development by single-cell RNA-sequencing. *bioRxiv*. https://doi.org/10.1101/478529 (2018).
13. Burmeister, Y. *et al.* ICOS controls the pool size of effector-memory and regulatory T cells. *The Journal of Immunology* **180,** 774–782 (2008).
14. Chen, X. & Oppenheim, J. J. Resolving the identity myth: key markers of functional CD4 + FoxP3 + regulatory T cells. *International Immunopharmacology* (2011).
15. Sunderkötter, C. *et al.* Subpopulations of mouse blood monocytes differ in maturation stage and inflammatory response. *The Journal of Immunology* **172,** 4410–4417 (2004).
16. Marcovecchio, P. M. *et al.* Scavenger receptor CD36 directs nonclassical monocyte patrolling along the endothelium during early atherogenesis. *Arteriosclerosis, thrombosis, and vascular biology* **37,** 2043–2052 (2017).
17. Loder, F. *et al.* B cell development in the spleen takes place in discrete steps and is determined by the quality of B cell receptor-derived signals. *Journal of Experimental Medicine* **190,** 75–89 (1999).
18. Miller, J. C. *et al.* Deciphering the transcriptional network of the dendritic cell lineage. *Nature Immunology* (2012).
19. Borges Da Silva, H. *et al.* Splenic macrophage subsets and their function during blood-borne infections. *Frontiers in Immunology* **6,** 480 (2015).
20. Tardif, M. R. *et al.* Secretion of S100A8, S100A9, and S100A12 by neutrophils involves reactive oxygen species and potassium efflux. *Journal of Immunology Research* (2015).
21. Fehniger, T. A. *et al.* Acquisition of Murine NK cell cytotoxicity requires the translation of a pre-existing pool of Granzyme B and Perforin mRNAs. *Immunity* **26,** 798–811 (2007).
22. Bendelac, A., Rivera, M. N., Park, S.-H. & Roark, J. H. Mouse CD1-Specific NK1 T Cells: Development, Specificity, and Function. *Annual Review of Immunology* **15,** 535–562 (1997).
23. Zhang, J. *et al.* Characterization of Siglec-H as a novel endocytic receptor expressed on murine plasmacytoid dendritic cell precursors. *Blood* **107,** 3600–3608 (2006).
24. Sawai, C. M. *et al.* Transcription factor Runx2 controls the development and migration of plasmacytoid dendritic cells. *Journal of Experimental Medicine* **210,** 2151–2159 (2013).
25. Castro, C. D. & Flajnik, M. F. Putting J Chain back on the map: How might its expression define plasma cell development? *The Journal of Immunology* **193,** 3248–3255 (2014).

26. Dutta, P. *et al.* Macrophages retain hematopoietic stem cells in the spleen via VCAM-1. *Journal of Experimental Medicine* **212,** 497–512 (2015).
27. Hobeika, E., Dautzenberg, M., Levit-Zerdoun, E., Pelanda, R. & Reth, M. Conditional selection of B cells in mice with an inducible B cell development. *Frontiers in Immunology* **9,** 1806 (2018).
28. Thomas, M. D., Srivastava, B. & Allman, D. Regulation of peripheral B cell maturation. *Cellular Immunology* **239,** 92–102 (2006).
29. Roncarolo, M.-G. & Gregori, S. Is FOXP3 a bona fide marker for human regulatory T cells? *European Journal of Immunology* **38,** 925–927 (2008).
30. Lopez, R., Regier, J., Cole, M. B., Jordan, M. I. & Yosef, N. Deep generative modeling for single-cell transcriptomics. *Nature Methods* **15,** 1053–1058 (2018).
31. Eraslan, G., Simon, L. M., Mircea, M., Mueller, N. S. & Theis, F. J. Single-cell RNA-seq denoising using a deep count autoencoder. *Nature Communications* **10,** 390 (2019).
32. Xu, C. *et al.* Harmonization and annotation of single-cell transcriptomics data with deep generative models. *bioRxiv,* 532895 (2019).
33. DeTomaso, D. *et al.* Functional interpretation of single cell similarity maps. *Nature Communications* (2019).
34. Marguerat, S. & Bähler, J. Coordinating genome expression with cell size. *Trends in Genetics* **28,** 560–565 (2012).
35. Stoeckius, M. *et al.* Simultaneous epitope and transcriptome measurement in single cells. *Nature Methods* (2017).
36. Yang, S. *et al.* Decontamination of ambient RNA in single-cell RNA-seq with DecontX. *Genome Biology* **21,** 57 (2020).
37. Blei, D. M. & Lafferty, J. D. A correlated topic model of Science. *The Annals of Applied Statistics* **1,** 17–35 (2007).
38. Dieng, A. B., Ruiz, F. J. R. & Blei, D. M. Topic Modeling in Embedding Spaces. *Transactions of the Association for Computational Linguistics* **8,** 439–453 (2020).
39. Kingma, D. P. & Welling, M. *Auto-Encoding variational Bayes* in *International Conference on Learning Representations* (2014).
40. Ioffe, S. & Szegedy, C. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv* (2015).
41. Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I. & Salakhutdinov, R. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research* **15,** 1929–1958 (2014).
42. Kucukelbir, A., Wang, Y. & Blei, D. M. *Evaluating Bayesian models with posterior dispersion indices* in *International Conference on Machine Learning* (2017).