Cell Reports Medicine, Volume *5*

Supplemental information

Cytometry masked autoencoder: An accurate and interpretable automated immunophenotyper

Jaesik Kim, Matei Ionita, Matthew Lee, Michelle L. McKeague, Ajinkya Pattekar, Mark M. Painter, Joost Wagenaar, Van Truong, Dylan T. Norton, Divij Mathew, Yonghyun Nam, Sokratis A. Apostolidis, Cynthia Clendenin, Patryk Orzechowski, Sang-Hyuk Jung, Jakob Woerner, Caroline A.G. Ittner, Alexandra P. Turner, Mika Esperanza, Thomas G. Dunn, Nilam S. Mangalmurti, John P. Reilly, Nuala J. Meyer, Carolyn S. Calfee, Kathleen D. Liu, Michael A. Matthy, Lamorna Brown Swigart, Ellen L. Burnham, Jeffrey McKeehan, Sheetal Gandotra, Derek W. Russel, Kevin W. Gibbs, Karl W. Thomas, Harsh Barot, Allison R. Greenplate, E. John Wherry, and Dokyoon Kim

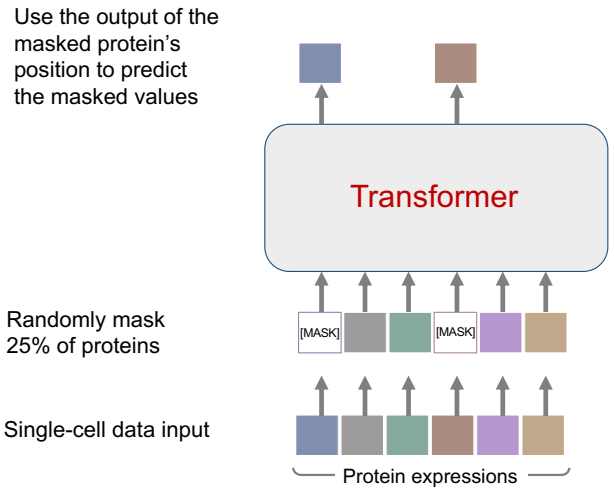


Figure S1. Overview of Masked Cytometry Modelling (*MCM*). Related to Figure 1.

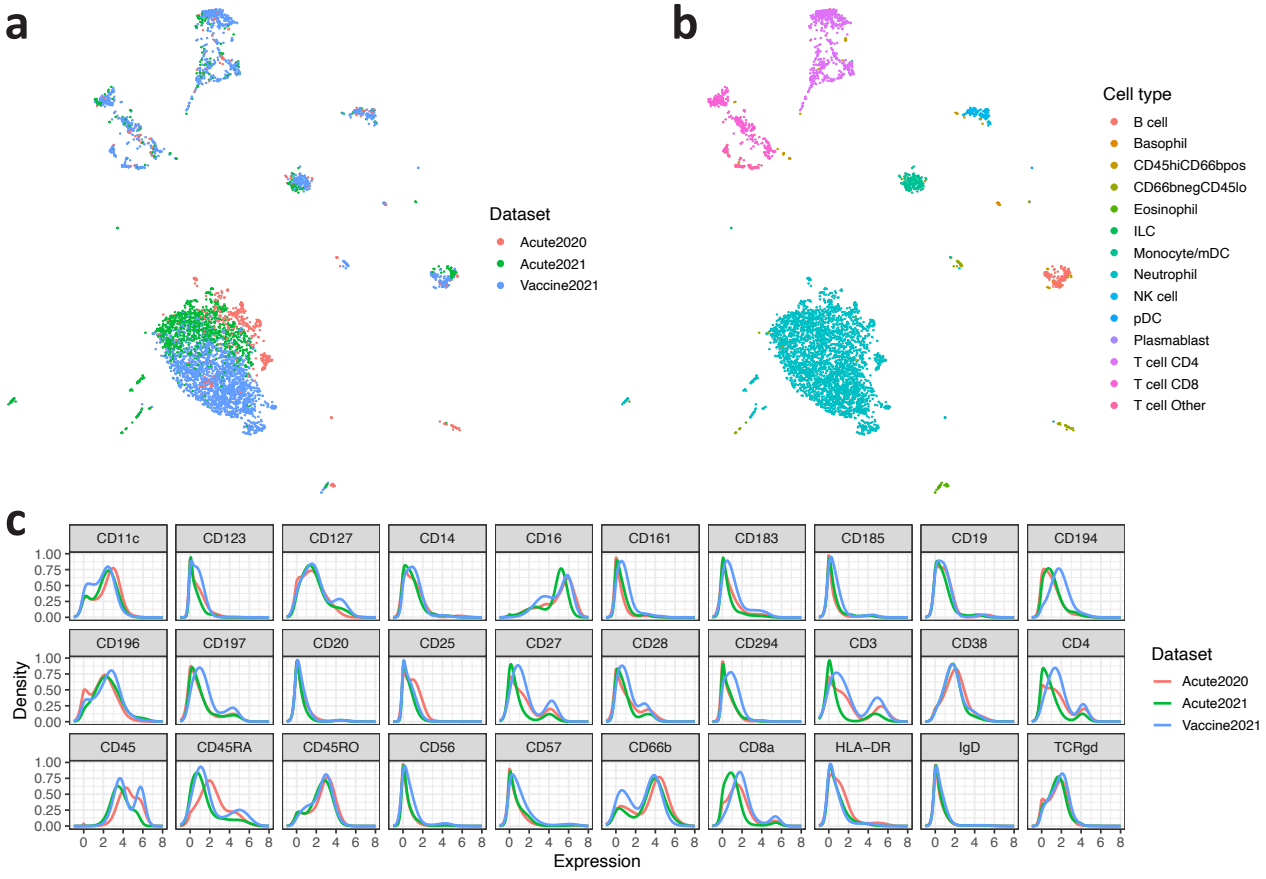


Figure S2. Batch effects between the three datasets. (a) UMAP projection of concatenated cells from all files, based on all 30 protein channels, color coded by dataset. (b) Same UMAP projection, color coded by cell type determined by manual gating. (c) Kernel density estimates of each protein channel, color coded by dataset. Related to Figure 2.

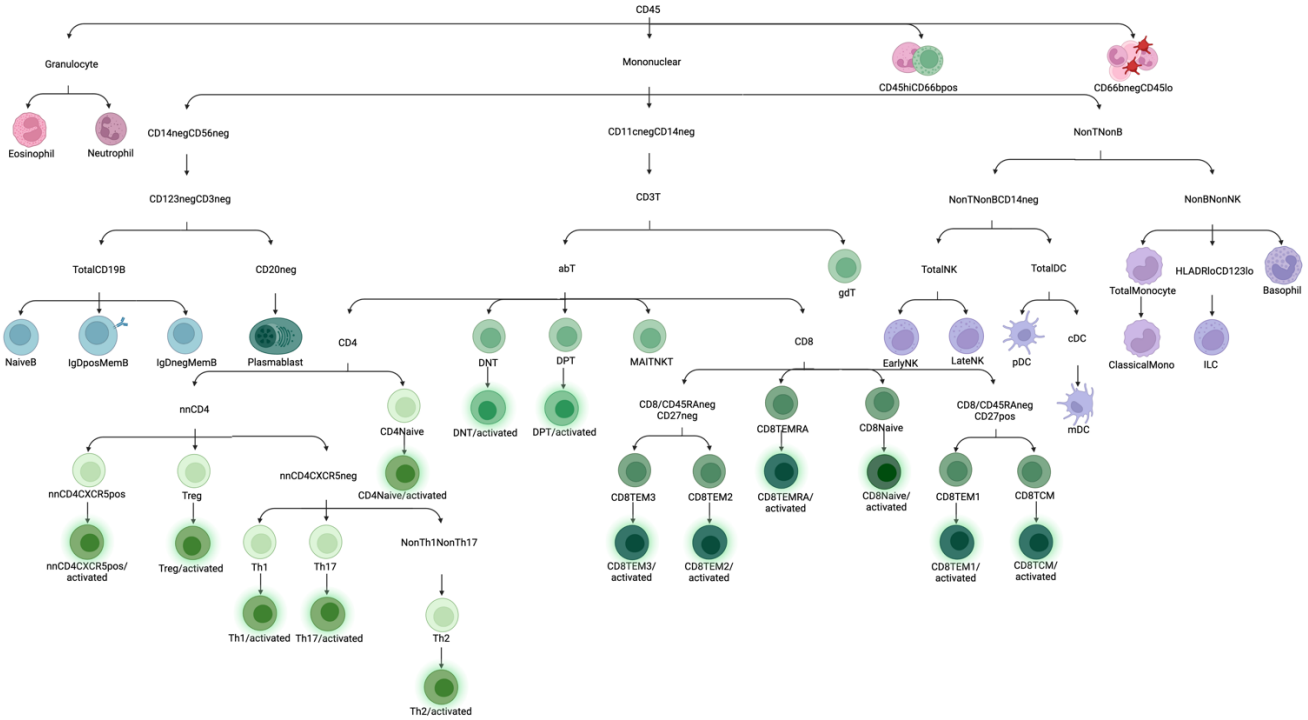


Figure S3. Standard gating strategies for 46 cell types. Related to Figure 2.

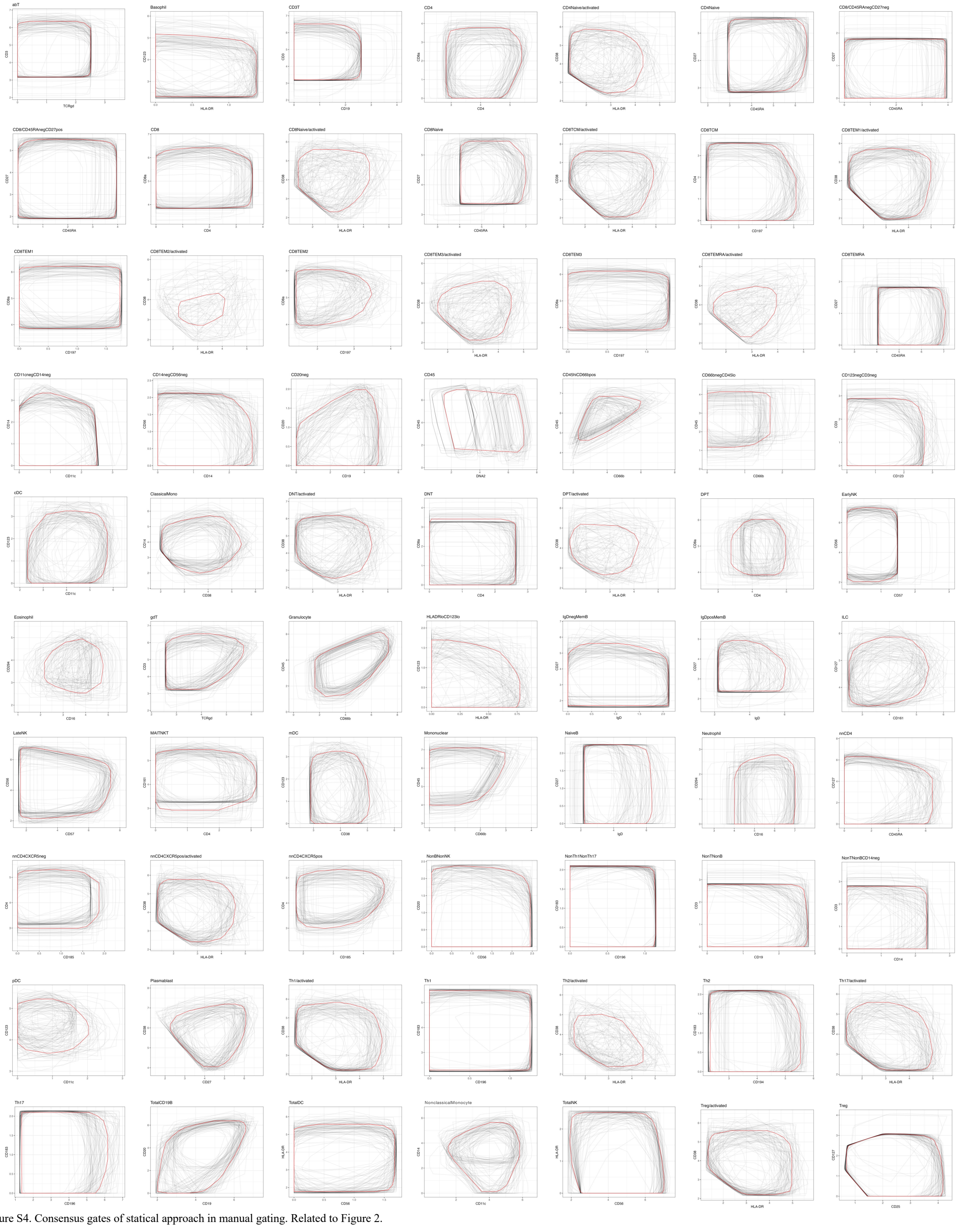


Figure S4. Consensus gates of statical approach in manual gating. Related to Figure 2.

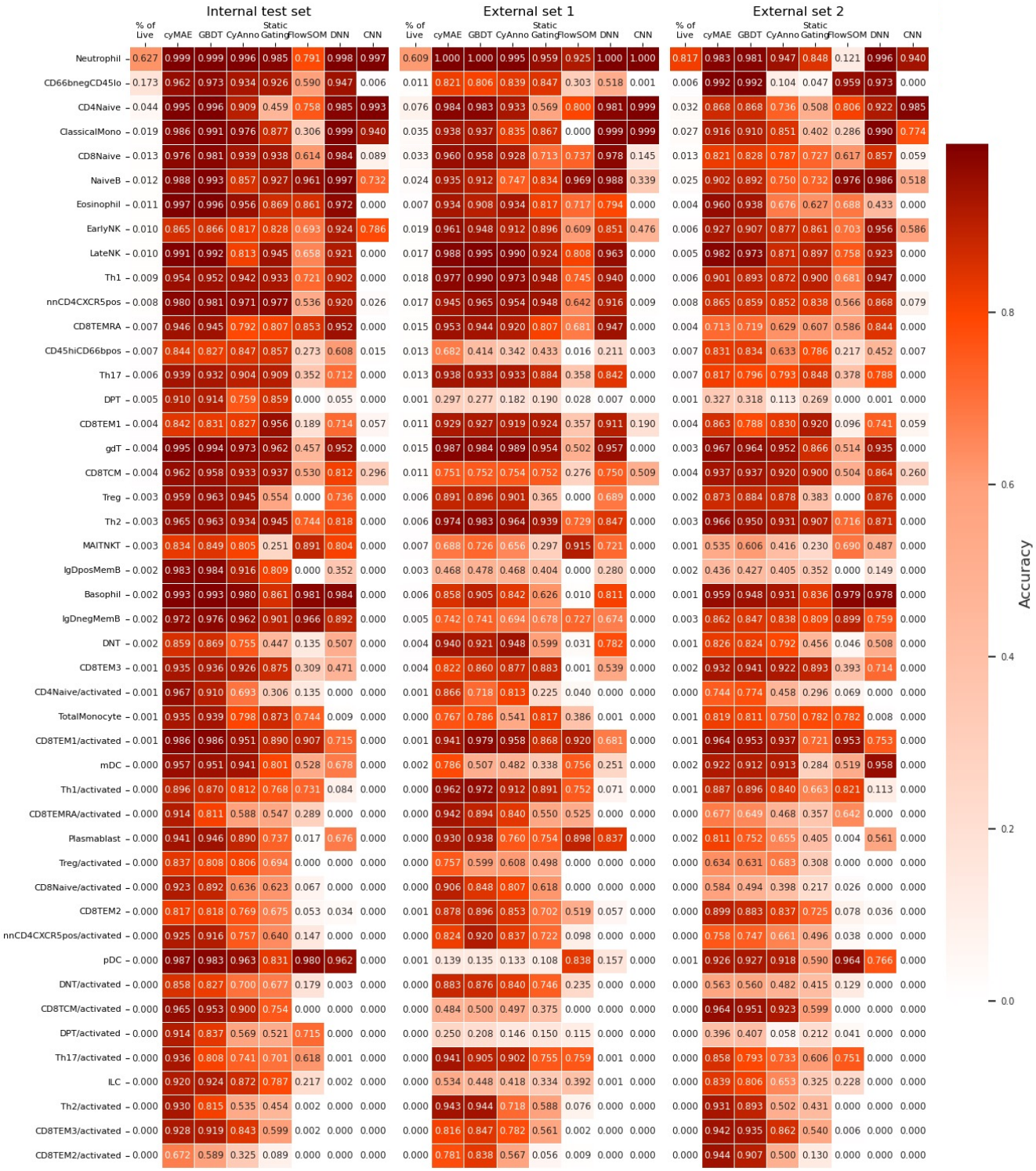


Figure S5. Full results of cell type prediction for each cell type. Related to Figure 2.

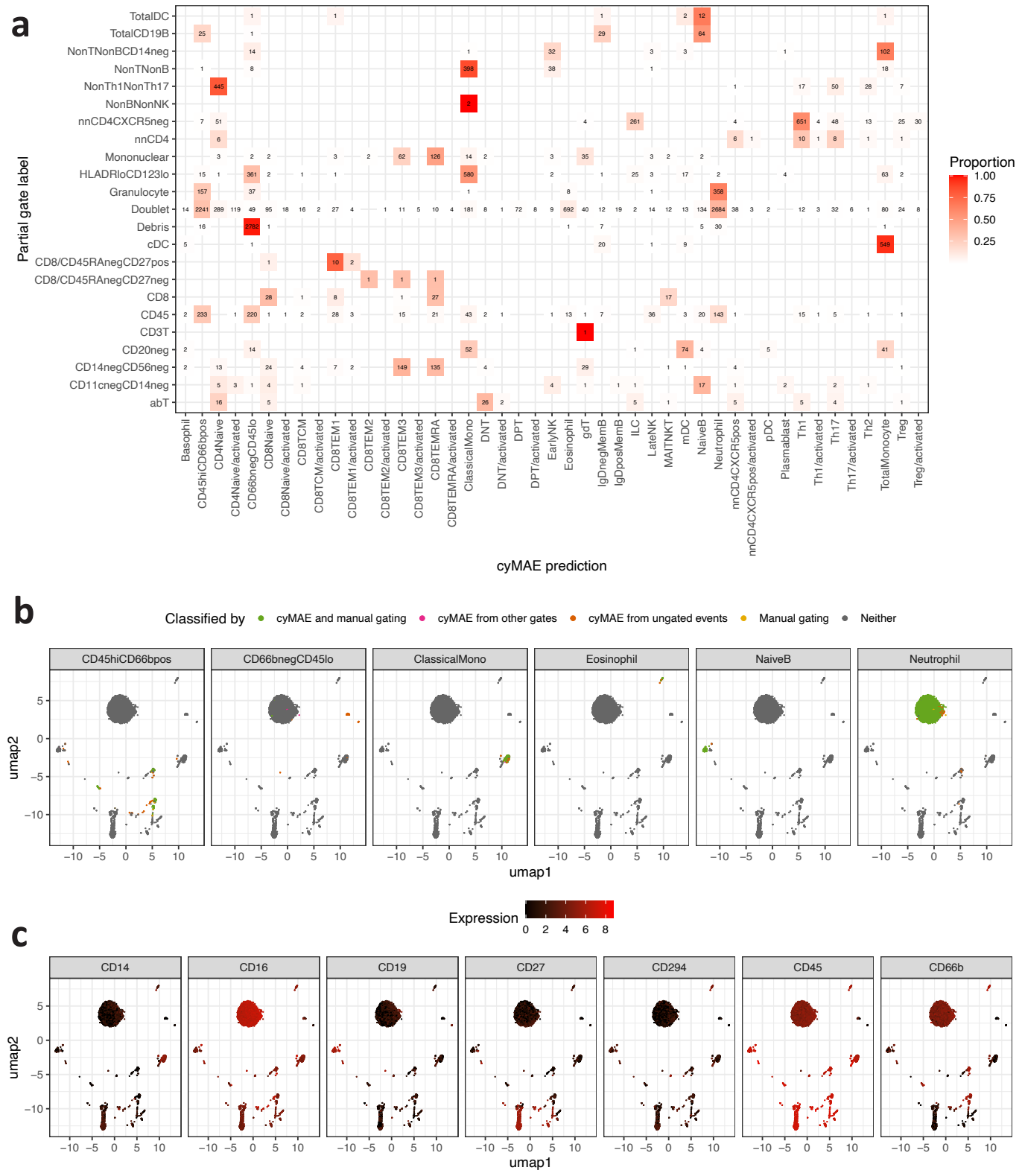


Figure S6. cyMAE performance on file containing both gated and ungated events. (a) Heatmap showing cyMAE predictions for events classified by manual gating as debris or doublets, or events which made it partly through the gating hierarchy before falling between the boundaries of downstream gates. Color shows proportion of events in each row which were assigned a given class by cyMAE. (b) UMAP projection of gated and ungated events. Each panel shows events classified as a given cell type by cyMAE, manual gating, both or neither. (c) Same UMAP projection, showing expression of selected proteins. Related to Figure 2.

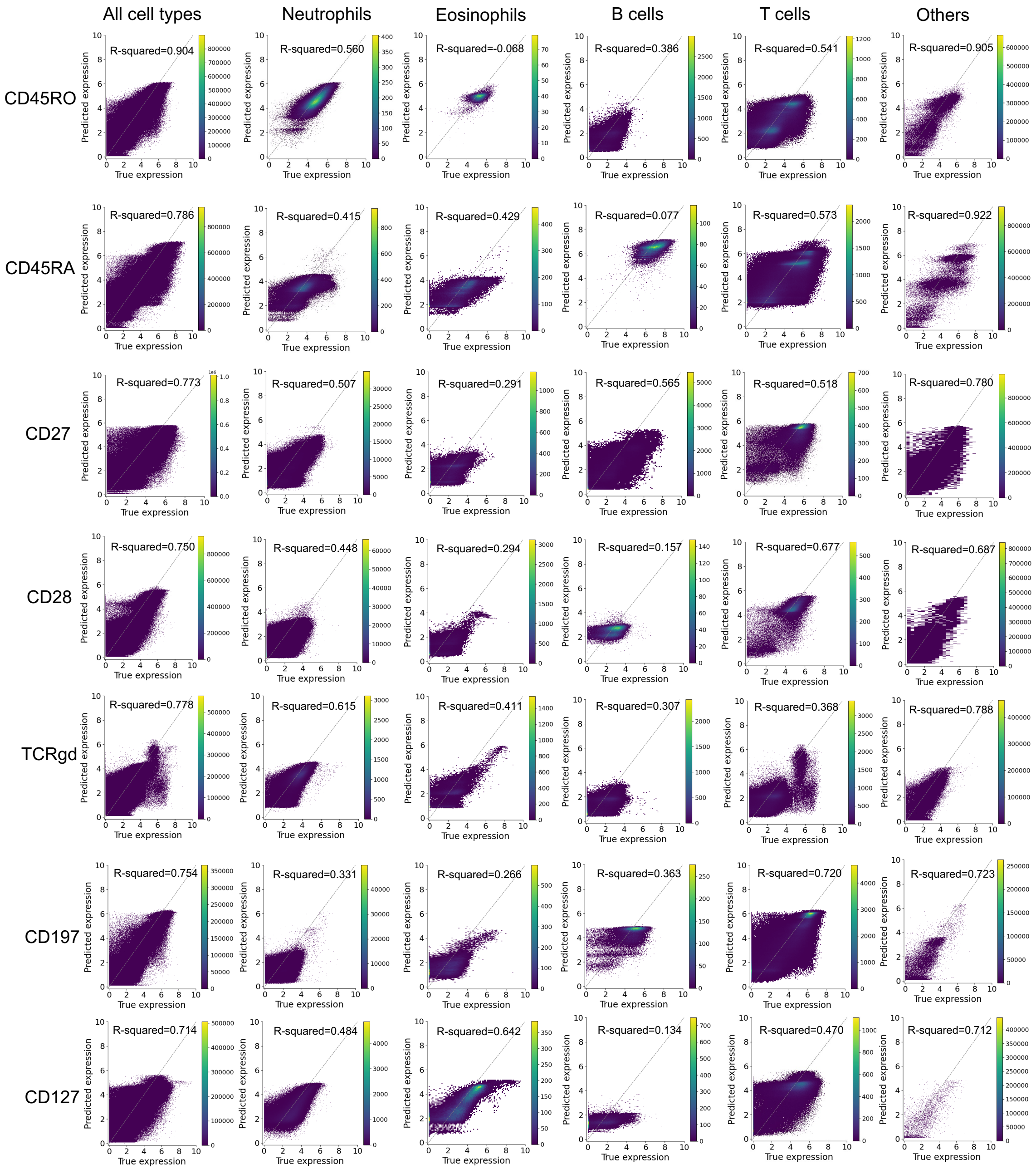


Figure S7. The true expression and imputed expression plots for the Acute2020 dataset. Related to Figure 3.

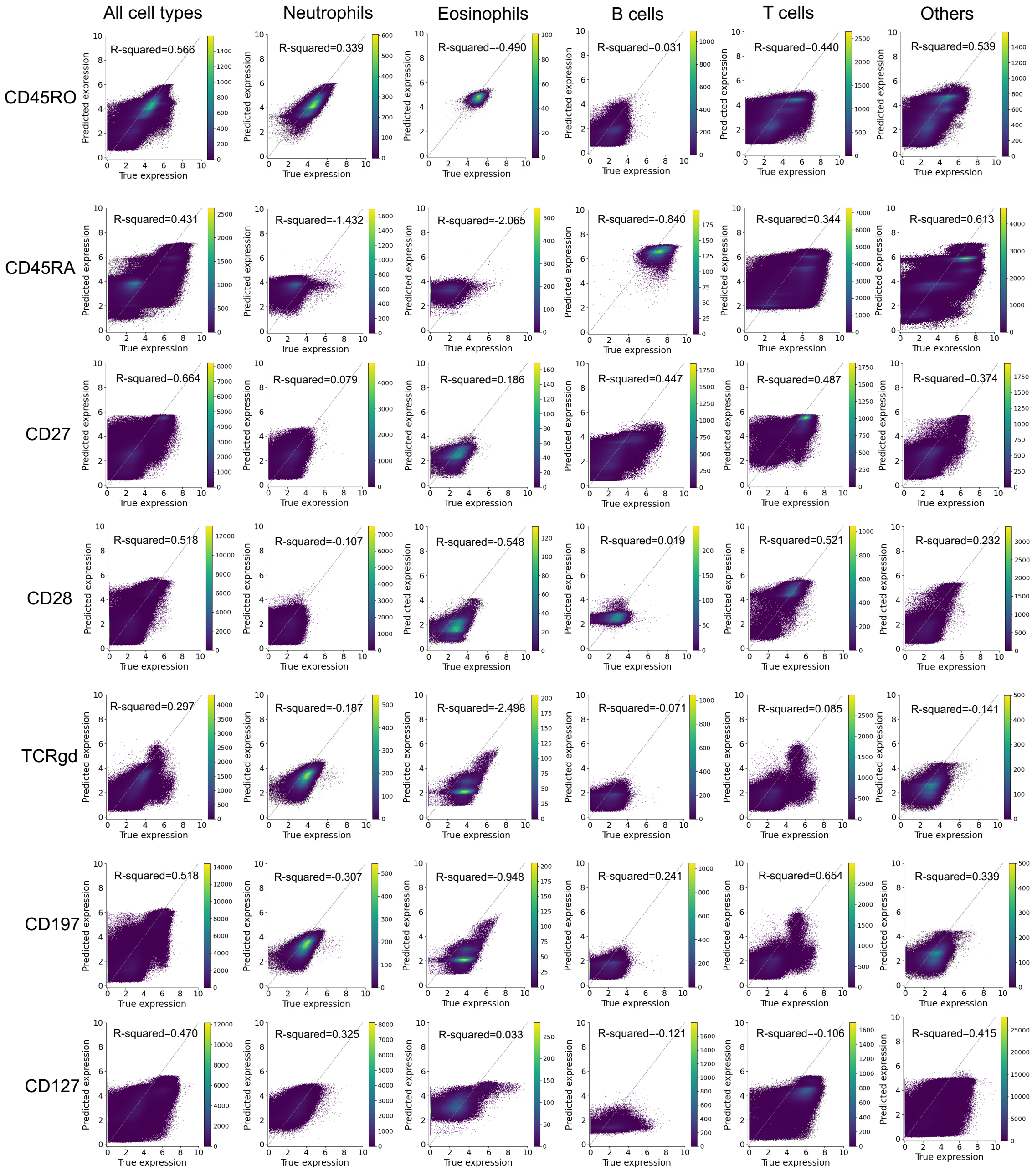


Figure S8. The true expression and imputed expression plots for the Vaccine dataset. Related to Figure 3.

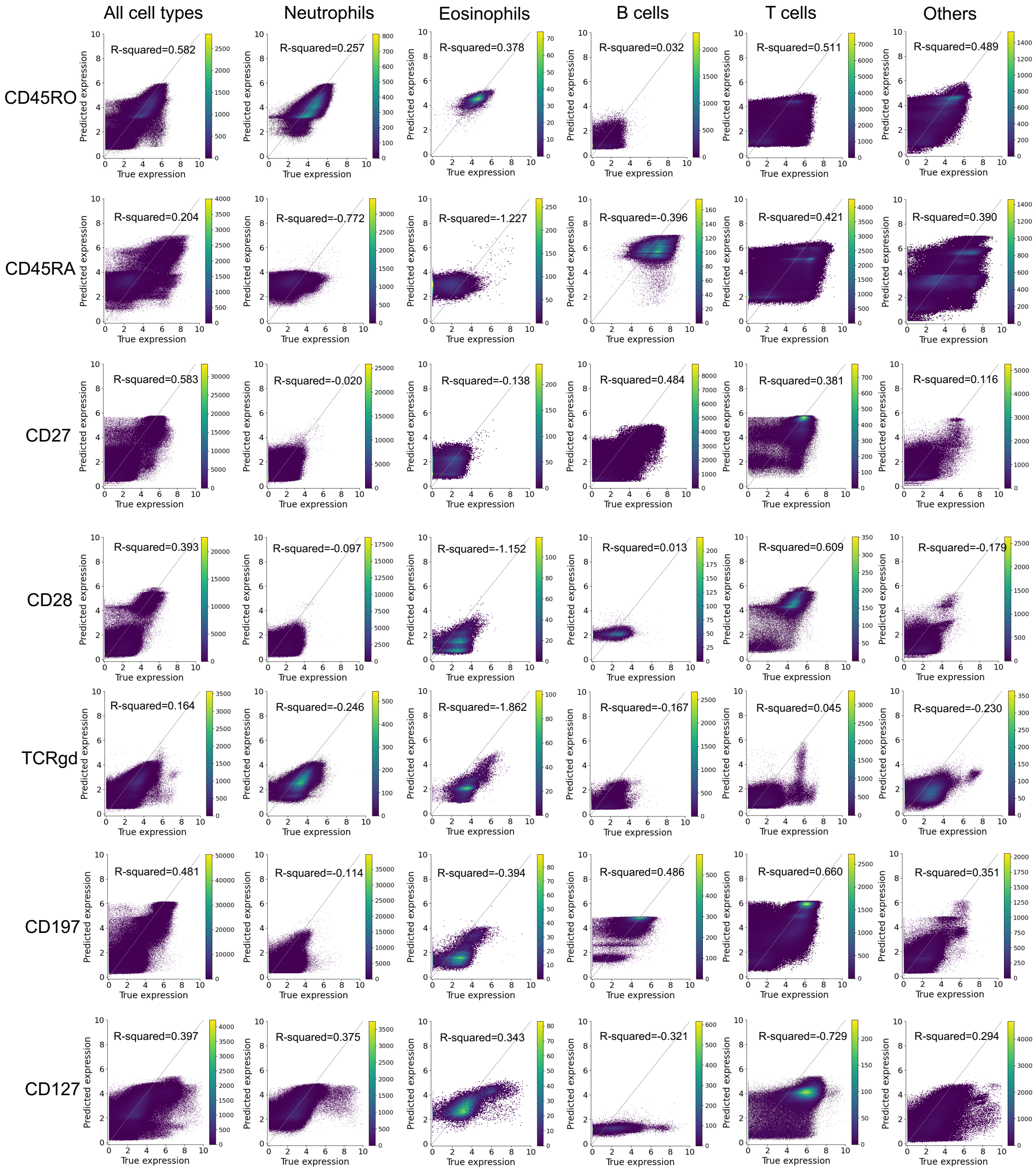


Figure S9. The true expression and imputed expression plots for the Acute2021 dataset. Related to Figure 3.

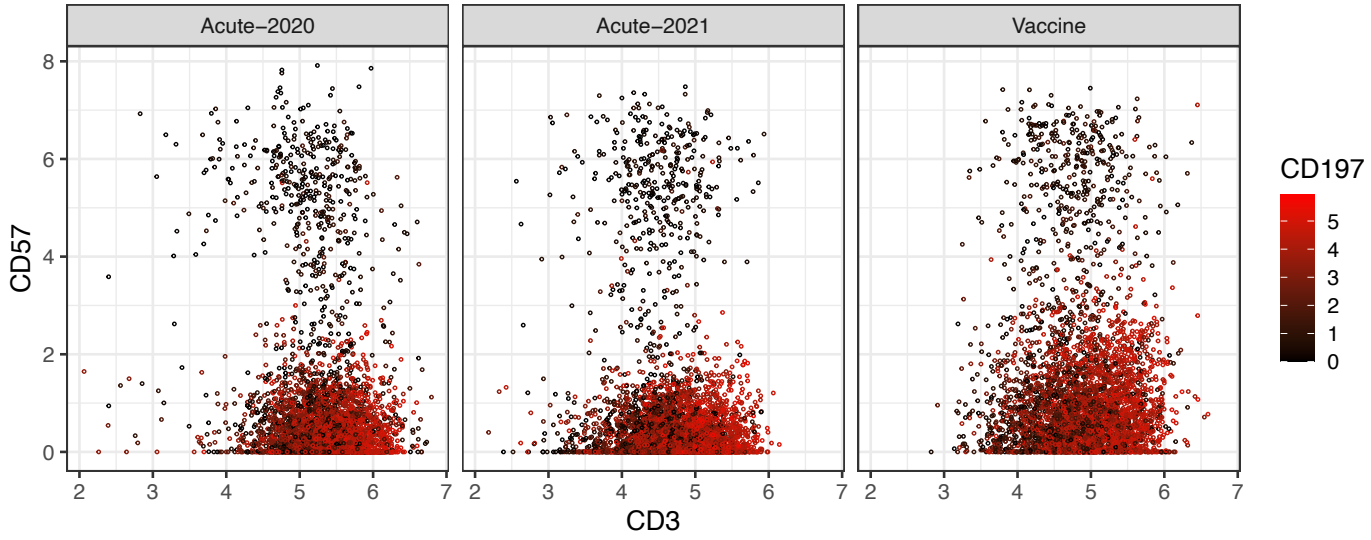


Figure S10. CD3 is positively correlated to CD197, and CD57 is negatively correlated to CD197 in T cells. Related to Figure 4.

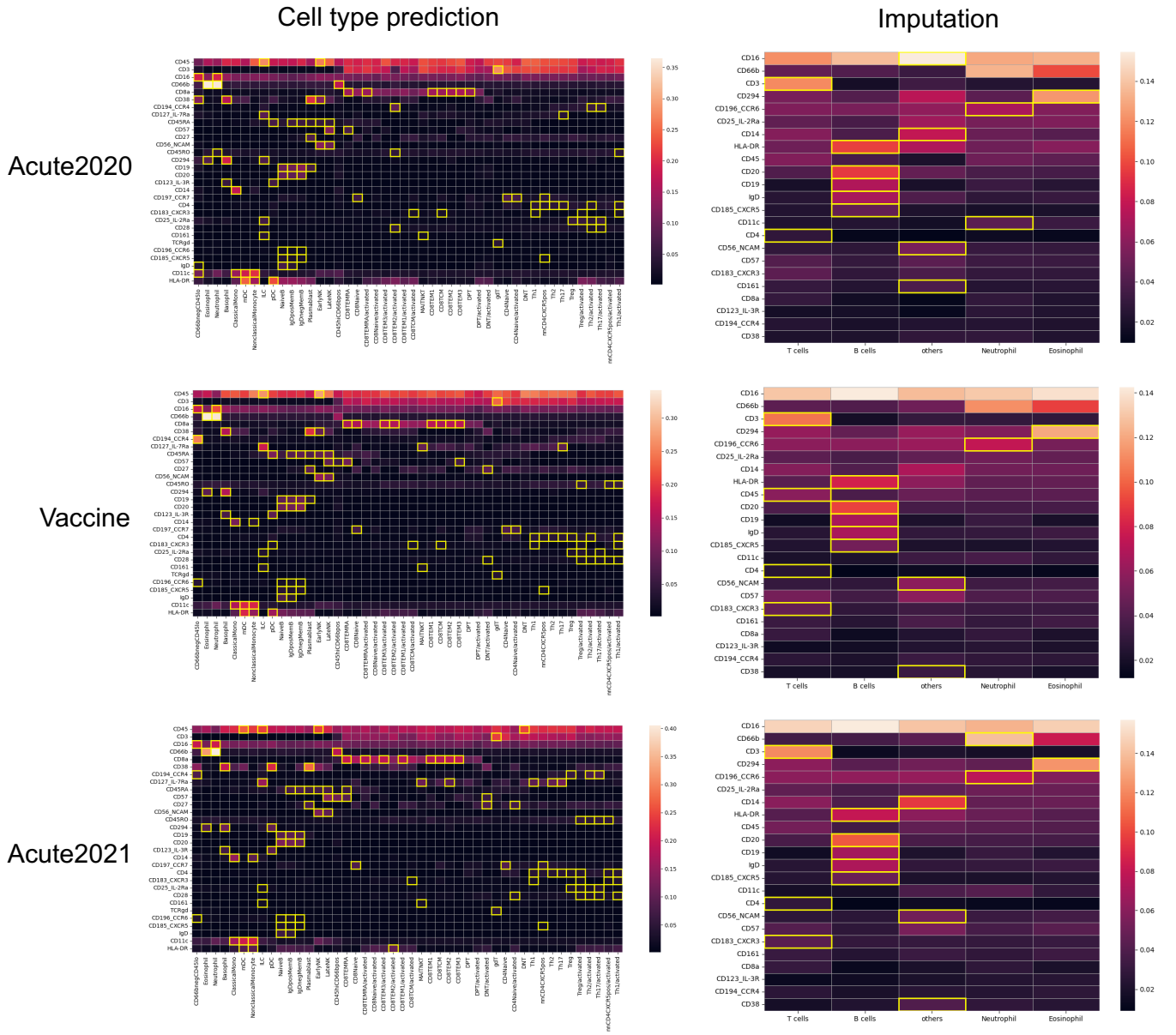


Figure S11. Attention scores of each cell type in the cell type classification and the imputation task. Related to Figure 4.

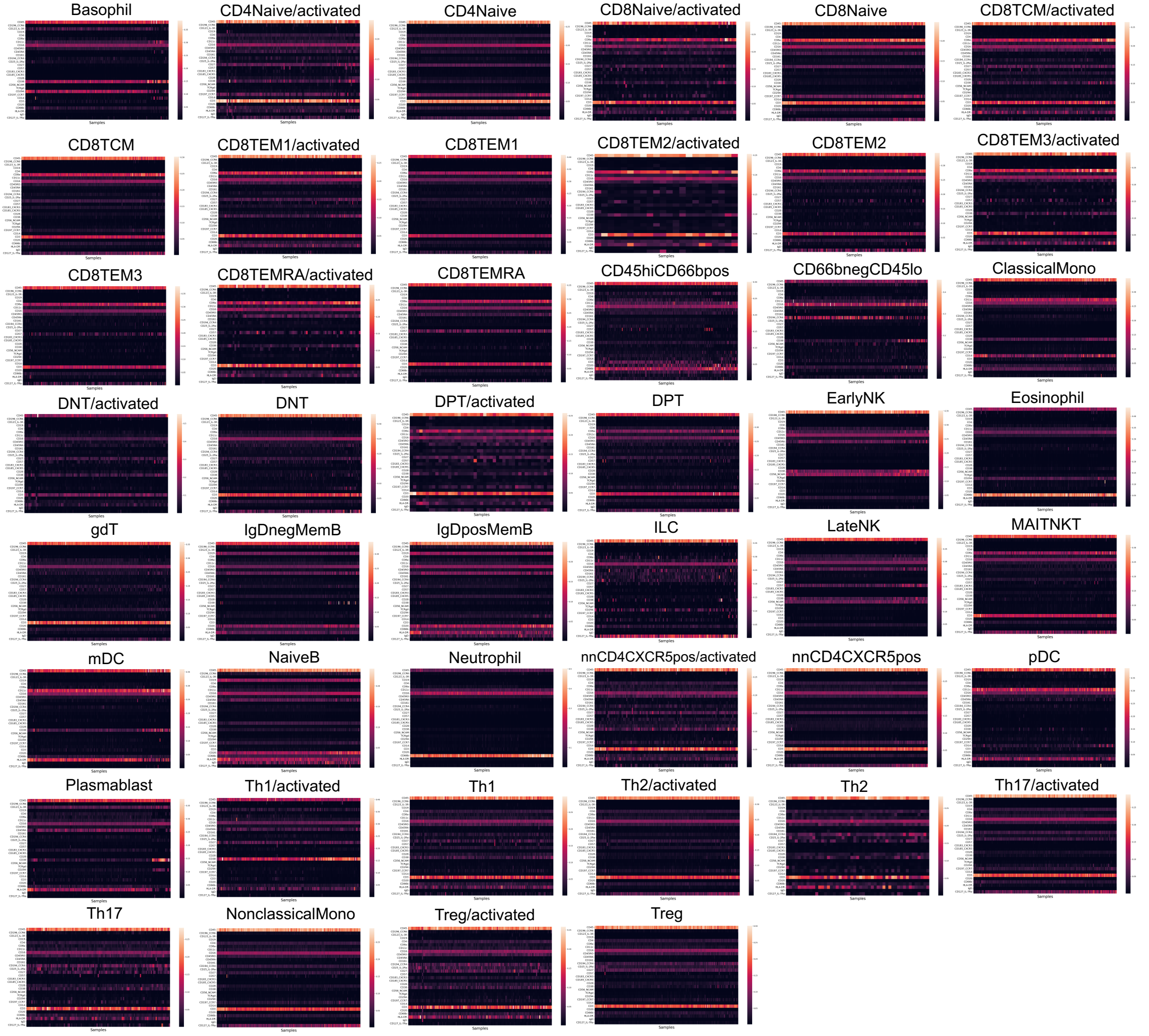


Figure S12. Attention scores of each cell type for entire samples in the cell type classification task. Related to Figure 4.

Figure S13. Attention scores of each cell type for entire samples in the imputation task. Related to Figure 4.

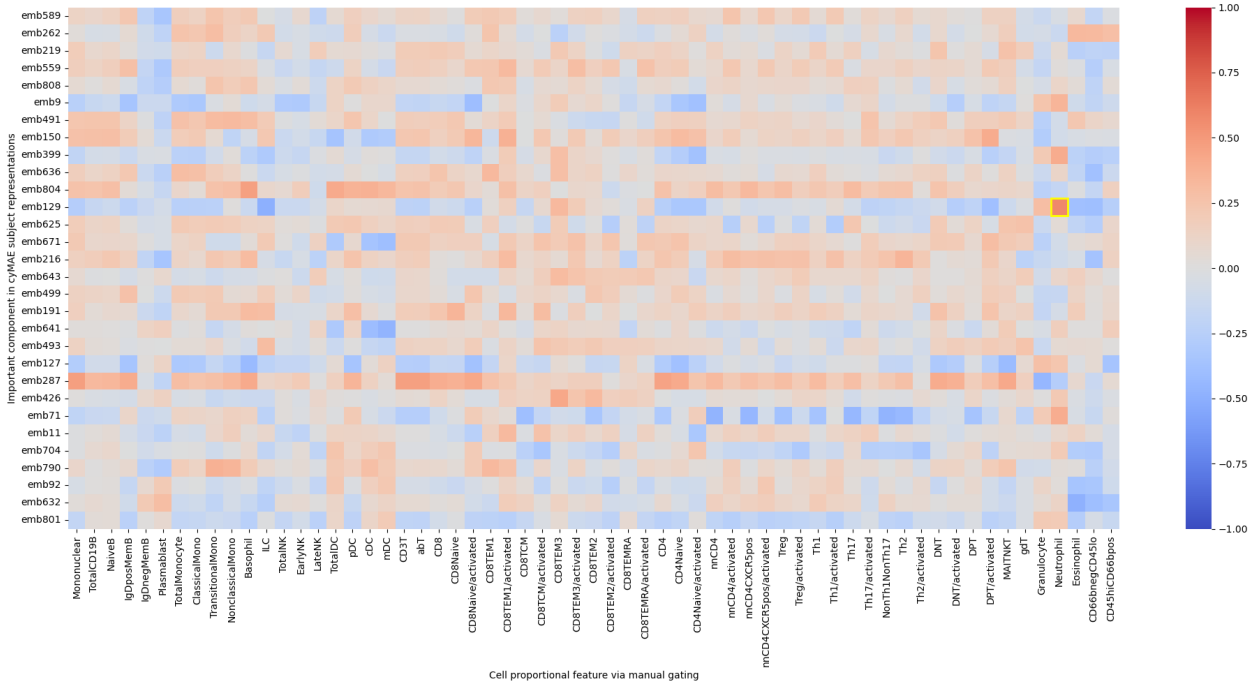
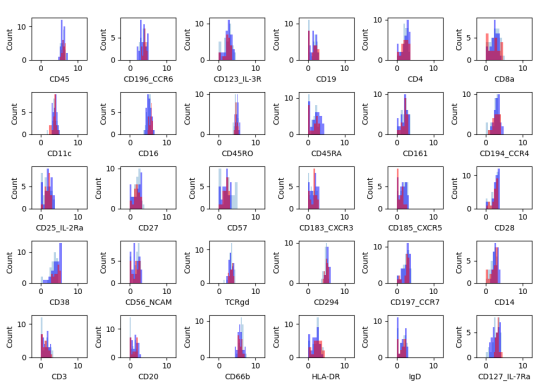
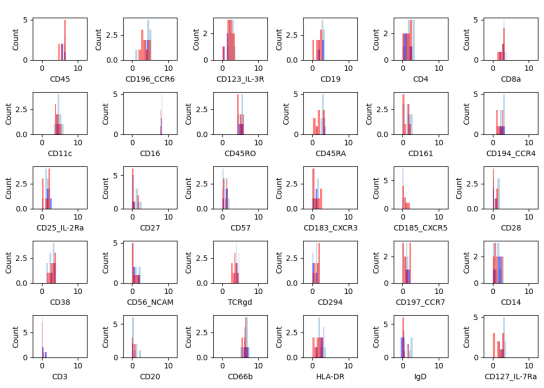


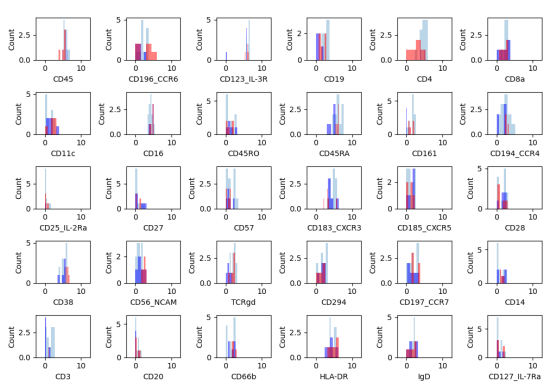
Figure S14. A heatmap shows the Pearson correlations between manual gated features (cell population proportion) and the selected components of the cyMAE subject representations for the COVID-19 pre- and post-treatment classification. Correlation above 0.5 is highlighted as yellow box. Related to Figure 6.



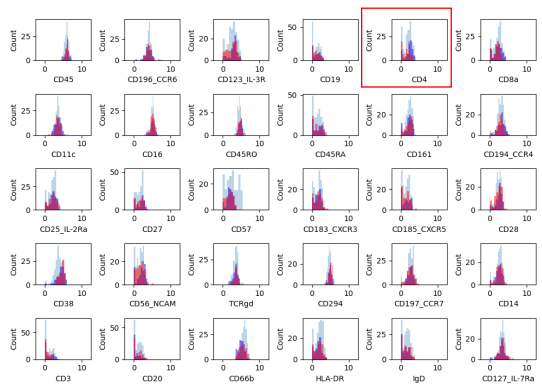
ROI 1 (Eosinophil)



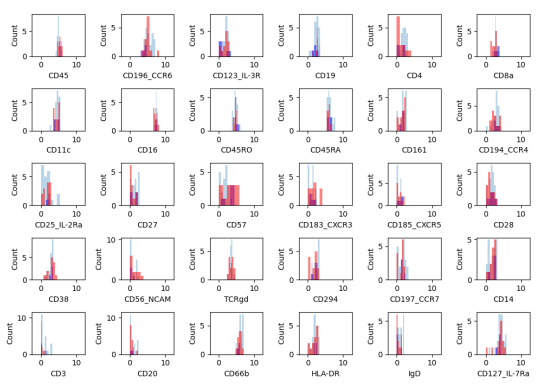
ROI 2 (Neutrophil)



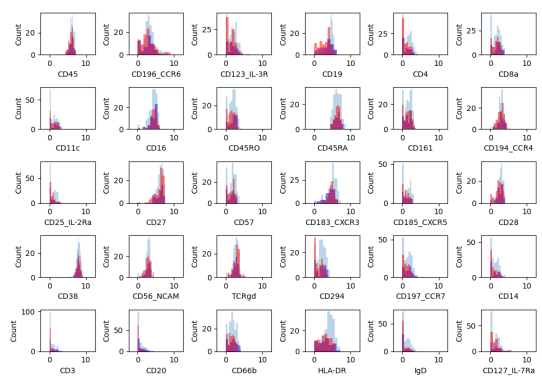
ROI 3 (pDC)



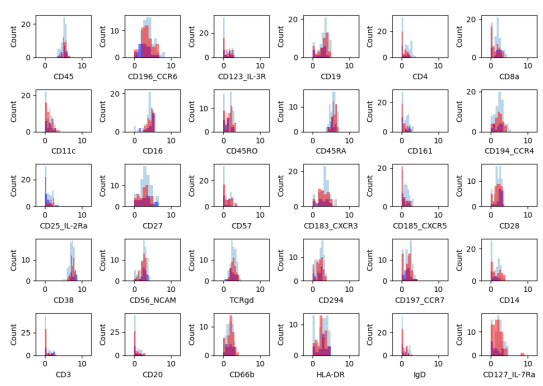
ROI 4 (Eosinophil/CD45hiCD66bpos)



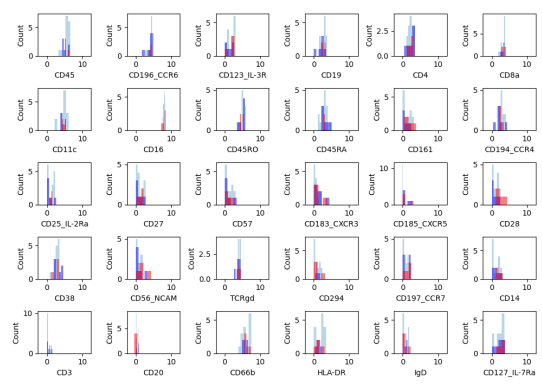
ROI 5 (Neutrophil)



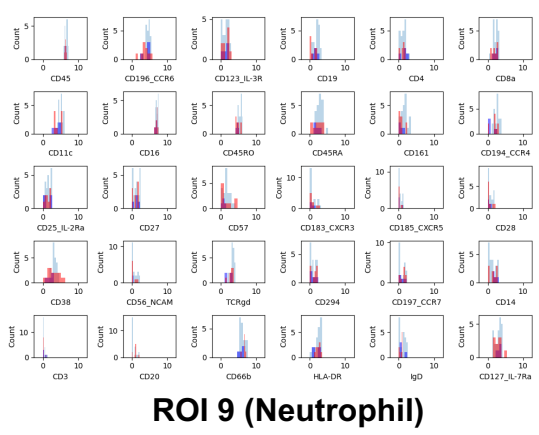
ROI 6 (Plasmablast/ CD66bnegCD45lo/IgDnegMemB/EarlyNK)



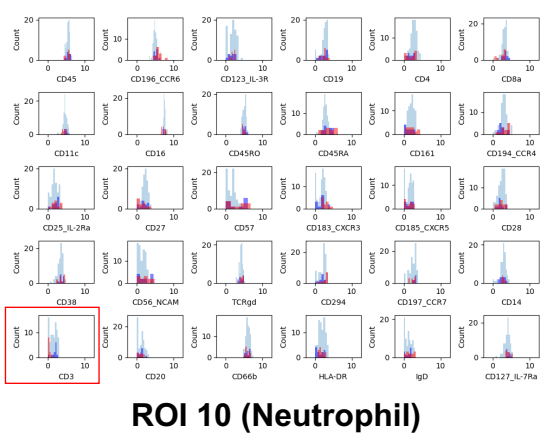
ROI 7 (CD66bnegCD45lo/ Plasmablast/IgDnegMemB)



ROI 8 (Neutrophil)



ROI 9 (Neutrophil)



ROI 10 (Neutrophil)

Figure S15. Histograms of marker expression of red starred cells (red histogram), blue starred cells (blue histogram), and background cells (sky-blue histogram) for each ROI in the COVID-19 pre- and post-treatment classification task. Significantly different marker expressions between pre-treatment associated cells and post-treatment associated cells are highlighted as a red box (in ROI 4 and 10) based on the Kolmogorov-Smirnov test with FDR p-value correction. Related to Figure 6.

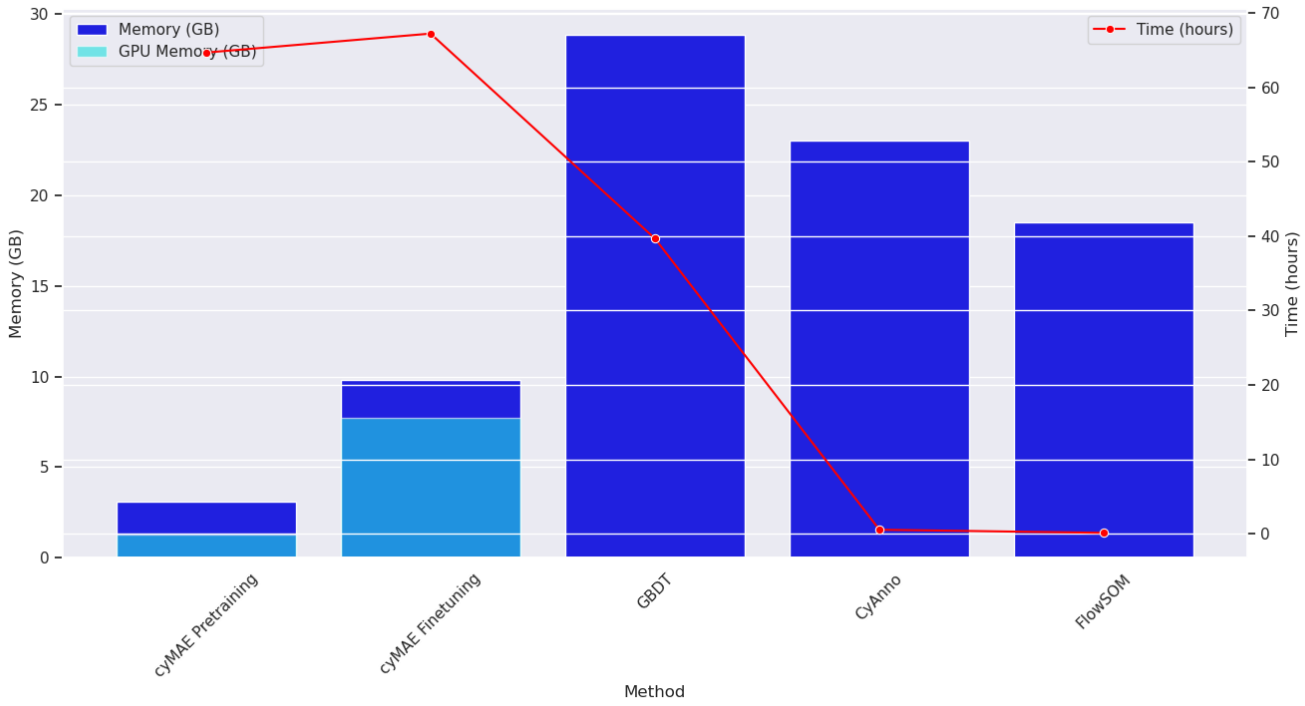


Figure S16. Comparison of training costs across methods in the cell type annotation. Memory usage is represented by the blue bar plot on the left axis, while runtime is indicated by the red line plot on the right axis. Different methods are compared in terms of memory (GB) and training time (hours). Related to STAR Methods.

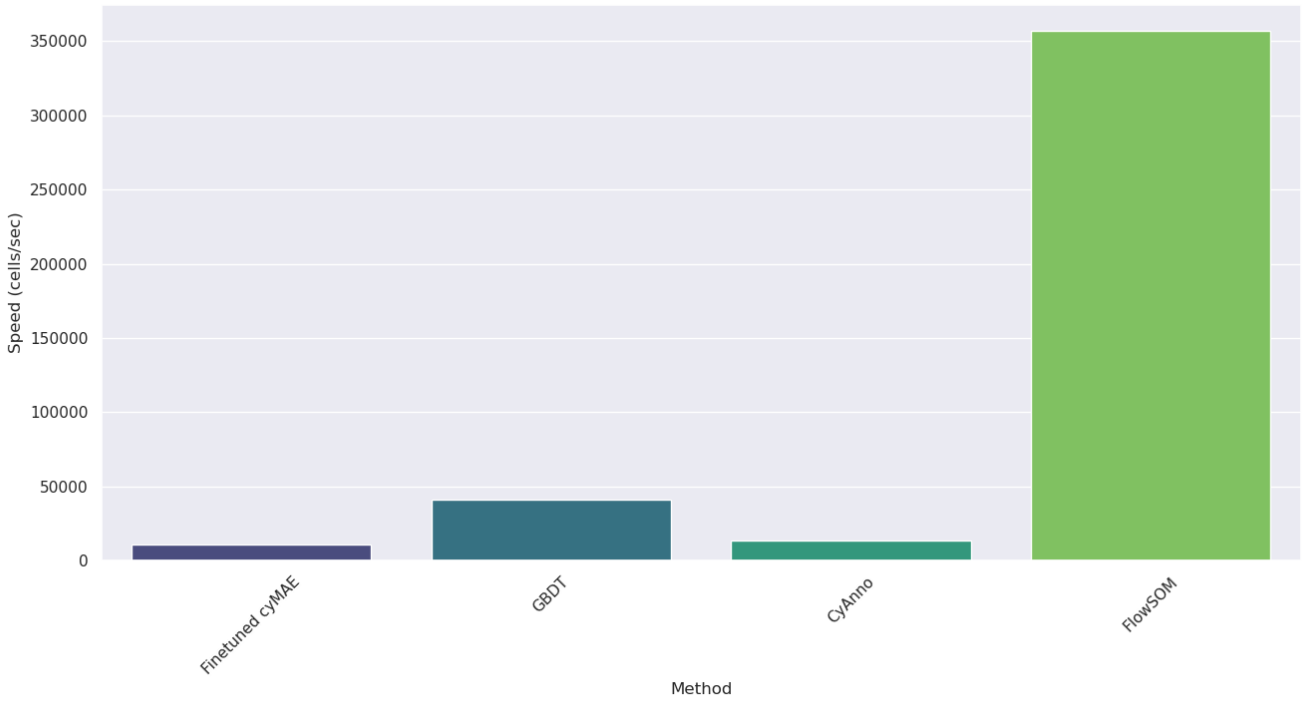


Figure S17. Comparison of inference speeds in the cell type annotation. The values show the amortized inference speeds based on processing three internal test set, external set 1, and external set 2. The speeds are aggregated and averaged to provide a practical and comprehensive comparison of different methods. Related to STAR Methods.

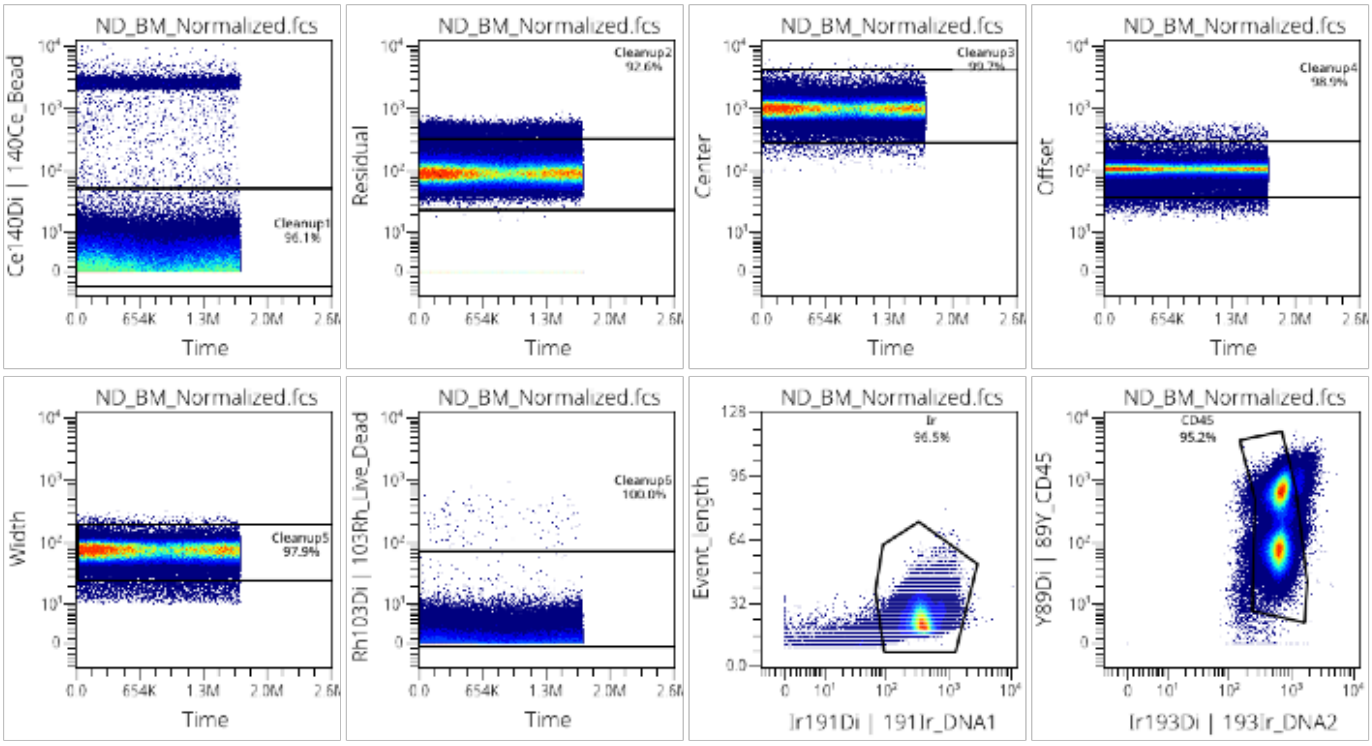


Figure S18. A standard cleanup procedure, which is a routine manual gating practice. fcs files were gated for beads, debris, doublets, and dead cells using the OMIQ platform. Related to STAR Methods.

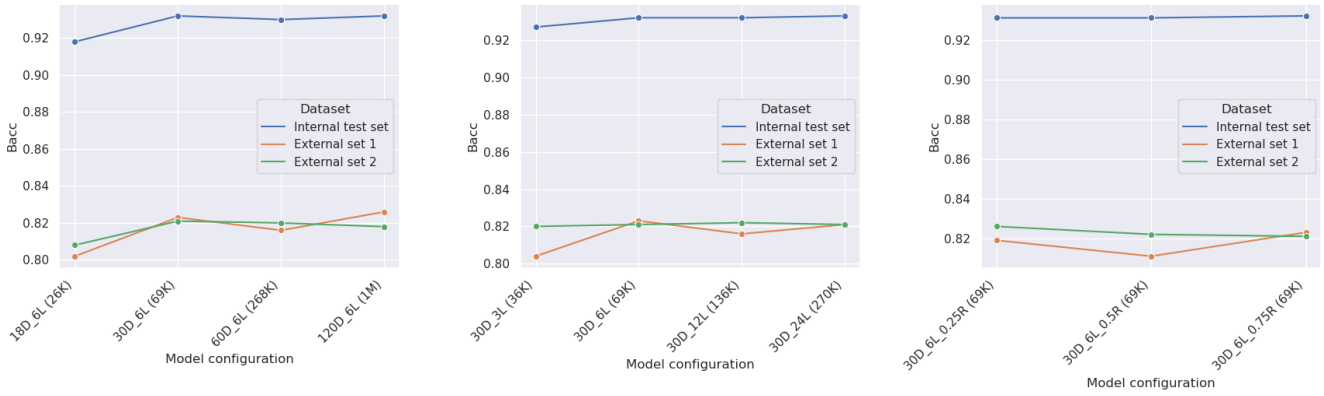


Figure S19. Exploration of model configurations based on the cell type annotation task. The model configuration is represented as "{Latent dimension}D_{number of layers}L_{masking ratio}R (Model size)". For example, "30D_6L_0.25R (69K)" denotes 30-dimensional latent representation for each marker, a 6-layer cyMAE architecture, pre-trained with 0.25 masking ratio, and a total of 69K parameters for the encoder and classifier after fine-tuning. Related to STAR Methods.

Table S1. Balanced accuracy comparison between the non-pre-trained and the pre-trained cyMAE in cell type annotation. Related to Figure 2.

	Internal test set (Bacc)	External set 1 (Bacc)	External set 2 (Bacc)
cyMAE from scratch	0.930	0.817	0.822
cyMAE with fine-tuning	0.931	0.819	0.826

Table S2. Accuracy and Balanced accuracy for cyMAE model trained on gated events and evaluated with our without ungated events present. Related to Figure 2.

	Acc	Bacc
With ungated, strict scoring	0.888	0.607
With ungated, lax scoring	0.955	0.813
Without ungated	0.989	0.897

Table S3. Full results of COVID-19 and healthy classification problem. Related to Figure 5.

Feature extraction methods		Predictors	Validation set (AUROC Mean ± Std.)	Test set (AUROC Mean ± Std.)
Manual gating		GBDT	**0.970 ± 0.091**	**0.975 ± 0.042**
		Logistic regression with L2 reg.	0.917 ± 0.047	0.938 ± 0.059
FlowSOM		GBDT	**0.930 ± 0.089**	**0.936 ± 0.096**
		Logistic regression with L2 reg.	0.875 ± 0.077	0.902 ± 0.090
CNN		CNN	0.633 ± 0.274	0.543 ± 0.256
cyMAE	Global mean pooling	GBDT	0.890 ± 0.135	0.888 ± 0.127
		Logistic regression with L2 reg.	0.923 ± 0.064	0.919 ± 0.077
	Global sum pooling	GBDT	0.909 ± 0.126	0.884 ± 0.102
		Logistic regression with L2 reg.	0.910 ± 0.098	0.859 ± 0.155
	Global max pooling	GBDT	0.993 ± 0.025	0.963 ± 0.068
		Logistic regression with L2 reg.	**0.993 ± 0.018**	**0.989 ± 0.033**
	Global min pooling	GBDT	0.942 ± 0.108	0.902 ± 0.127
		Logistic regression with L2 reg.	**0.980 ± 0.038**	**0.982 ± 0.038**

* reg. stands for regularization, Std. stands for standard deviation, and AUROC stands for Area Under the Receiver Operating Characteristic curve.

Table S4. Full results of Secondary immune response against COVID-19 prediction problem. Related to Figure 5.

Feature extraction methods		Predictors	Validation set (AUROC Mean ± Std.)	Test set (AUROC Mean ± Std.)
Manual gating		GBDT	**0.735 ± 0.129**	**0.641 ± 0.154**
		Logistic regression with L2 reg.	0.456 ± 0.171	0.446 ± 0.140
FlowSOM		GBDT	**0.622 ± 0.132**	**0.579 ± 0.151**
		Logistic regression with L2 reg.	0.577 ± 0.162	0.520 ± 0.167
CNN		CNN	0.585 ± 0.157	0.520 ± 0.163
cyMAE	Global mean pooling	GBDT	0.605 ± 0.166	0.535 ± 0.136
		Logistic regression with L2 reg.	0.626 ± 0.134	0.588 ± 0.125
	Global sum pooling	GBDT	0.613 ± 0.169	0.560 ± 0.132
		Logistic regression with L2 reg.	0.643 ± 0.126	0.607 ± 0.138
	Global max pooling	GBDT	0.635 ± 0.146	0.584 ± 0.153
		Logistic regression with L2 reg.	**0.608 ± 0.119**	**0.625 ± 0.154**
	Global min pooling	GBDT	0.616 ± 0.167	0.559 ± 0.179
		Logistic regression with L2 reg.	**0.674 ± 0.132**	**0.668 ± 0.157**

* reg. stands for regularization, Std. stands for standard deviation, and AUROC stands for Area Under the Receiver Operating Characteristic curve.

Table S5. Full results of COVID-19 pre- and post-treatment classification problem. Related to Figure 5.

Feature extraction methods		Predictors	Validation set (AUROC Mean ± Std.)	Test set (AUROC Mean ± Std.)
Manual gating		GBDT	**0.865 ± 0.126**	**0.796 ± 0.124**
		Logistic regression with L2 reg.	0.621 ± 0.120	0.615 ± 0.177
FlowSOM		GBDT	**0.885 ± 0.106**	**0.859 ± 0.112**
		Logistic regression with L2 reg.	0.592 ± 0.134	0.579 ± 0.170
CNN		CNN	0.591 ± 0.173	0.531 ± 0.201
cyMAE	Global mean pooling	GBDT	0.692 ± 0.171	0.610 ± 0.178
		Logistic regression with L2 reg.	0.740 ± 0.158	0.714 ± 0.155
	Global sum pooling	GBDT	0.663 ± 0.189	0.651 ± 0.149
		Logistic regression with L2 reg.	0.690 ± 0.180	0.676 ± 0.155

Global max pooling	GBDT	0.750 ± 0.160	0.718 ± 0.152
	Logistic regression with L2 reg.	**0.895 ± 0.103**	**0.887 ± 0.114**
Global min pooling	GBDT	0.818 ± 0.139	0.821 ± 0.138
	Logistic regression with L2 reg.	**0.919 ± 0.097**	**0.858 ± 0.121**

* reg. stands for regularization, Std. stands for standard deviation, and AUROC stands for Area Under the Receiver Operating Characteristic curve.