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 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.

% of		Inte	ernal	test	set			% of		E>	terna	al set	1			% of		Ex	terna	al set	2			
Live	CYMAE	GBDT	CyAnno	Gating	lowSOM		CNN	Live	cyMAE	GBDT	CyAnno	Gating	lowSOM	DNN	CNN	Live	CYMAE	GBDT	CyAnno	Gating	lowSOM		CNN	
Neutrophil – 0.627	0.999	0.999	0.996	0.985	0.791	0.998	0.997	0.609	1.000	1.000	0.995	0.959	0.925	1.000	1.000	0.817	0.983	0.981	0.947	0.848	0.121	0.996	0.940	
CD66bnegCD45lo - 0.173	0.962	0.973	0.934	0.926	0.590	0.947	0.006	0.011	0.821	0.806	0.839	0.847	0.303	0.518	0.001	0.006	0.992	0.992	0.104	0.047	0.959	0.973	0.000	
CD4Naive - 0.044	0.995	0.996	0.909	0.459	0.758	0.985	0.993	0.076	0.984	0.983	0.933	0.569	0.800	0.981	0.999	0.032	0.868	0.868	0.736	0.508	0.806	0.922	0.985	
ClassicalMono - 0.019	0.986	0.991	0.976	0.877	0.306	0.999	0.940	0.035	0.938	0.937	0.835	0.867	0.000	0.999	0.999	0.027	0.916	0.910	0.851	0.402	0.286	0.990	0.774	
CD8Naive - 0.013	0.976	0.981	0.939	0.938	0.614	0.984	0.089	0.033	0.960	0.958	0.928	0.713	0.737	0.978	0.145	0.013	0.821	0.828	0.787	0.727	0.617	0.857	0.059	
NaiveB - 0.012	0.988	0.993	0.857	0.927	0.961	0.997	0.732	0.024	0.935	0.912	0.747	0.834	0.969	0.988	0.339	0.025	0.902	0.892	0.750	0.732	0.976	0.986	0.518	
Eosinophil - 0.011	0.997	0.996	0.956	0.869	0.861	0.972	0.000	0.007	0.934	0.908	0.934	0.817	0.717	0.794	0.000	0.004	0.960	0.938	0.676	0.627	0.688	0.433	0.000	
EarlyNK = 0.010	0.865	0 866	0.817	0.828	0.693	0 924	0 786	0.019	0.961	0.948	0.912	0.896	0.609	0.851	0.476	0.006	0.927	0 907	0.877	0.861	0 703	0.956	0.586	
LateNK - 0.010	0.991	0.992	0.913	0.945	0.659	0.021	0.000	0.017	0.000	0.995	0.990	0.924	0.909	0.963	0.000	0.005	0.092	0.073	0.971	0.997	0.759	0.023	0.000	
Th1 0.000	0.991	0.992	0.013	0.945	0.050	0.921	0.000	0.017	0.900	0.995	0.990	0.924	0.000	0.905	0.000	0.005	0.902	0.975	0.071	0.097	0.750	0.923	0.000	
IIII - 0.009	0.954	0.952	0.942	0.955	0.721	0.902	0.000	0.016	0.977	0.990	0.975	0.946	0.745	0.940	0.000	0.000	0.901	0.095	0.072	0.900	0.001	0.947	0.000	
nncD4CXCR5pos = 0.008	0.980	0.981	0.971	0.977	0.536	0.920	0.026	0.017	0.945	0.965	0.954	0.948	0.642	0.916	0.009	0.008	0.865	0.859	0.852	0.838	0.566	0.868	0.079	- 0.8
CD8TEMRA - 0.007	0.946	0.945	0.792	0.807	0.853	0.952	0.000	0.015	0.953	0.944	0.920	0.807	0.681	0.947	0.000	0.004	0.713	0.719	0.629	0.607	0.586	0.844	0.000	
CD45hiCD66bpos – 0.007	0.844	0.827	0.847	0.857	0.273	0.608	0.015	0.013	0.682	0.414	0.342	0.433	0.016	0.211	0.003	0.007	0.831	0.834	0.633	0.786	0.217	0.452	0.007	
Th17 - 0.006	0.939	0.932	0.904	0.909	0.352	0.712	0.000	0.013	0.938	0.933	0.933	0.884	0.358	0.842	0.000	0.007	0.817	0.796	0.793	0.848	0.378	0.788	0.000	
DPT - 0.005	0.910	0.914	0.759	0.859	0.000	0.055	0.000	0.001	0.297	0.277	0.182	0.190	0.028	0.007	0.000	0.001	0.327	0.318	0.113	0.269	0.000	0.001	0.000	
CD8TEM1 - 0.004	0.842	0.831	0.827	0.956	0.189	0.714	0.057	0.011	0.929	0.927	0.919	0.924	0.357	0.911	0.190	0.004	0.863	0.788	0.830	0.920	0.096	0.741	0.059	
gdT - 0.004	0.995	0.994	0.973	0.962	0.457	0.952	0.000	0.015	0.987	0.984	0.989	0.954	0.502	0.957	0.000	0.003	0.967	0.964	0.952	0.866	0.514	0.935	0.000	
CD8TCM - 0.004	0.962	0.958	0.933	0.937	0.530		0.296	0.011		0.752	0.754		0.276	0.750	0.509	0.004	0.937	0.937	0.920	0.900	0.504	0.864	0.260	- 0.6
Treg = 0.003	0.959	0.963	0.945	0.554	0.000		0.000	0.006	0.891	0.896	0.901	0.365	0.000	0.689	0.000	0.002	0.873	0.884	0.878	0.383	0.000	0.876	0.000	- 0.6
Th2 - 0.003	0.965	0.963	0.934	0.945	0.744	0.818	0.000	0.006	0.974	0.983	0.964	0.939	0.729	0.847	0.000	0.003	0.966	0.950	0.931	0.907	0.716	0.871	0.000	
MAITNKT - 0.003	0.834	0.849	0.805	0.251	0.891	0.804	0.000	0.007	0.688	0.726	0.656	0.297	0.915	0.721	0.000	0.001	0.535	0.606	0.416	0.230	0.690	0.487	0.000	2
lgDposMemB – 0.002	0.983	0.984	0.916	0.809	0.000	0.352	0.000	0.003	0.468	0.478	0.468	0.404	0.000	0.280	0.000	0.002	0.436	0.427	0.405	0.352	0.000	0.149	0.000	Irac
Basophil - 0.002	0.993	0.993	0.980	0.861	0.981	0.984	0.000	0.006	0.858	0.905	0.842	0.626	0.010	0.811	0.000	0.001	0.959	0.948	0.931	0.836	0.979	0.978	0.000	CCL
lgDnegMemB – 0.002	0.972	0.976	0.962	0.901	0.966	0.892	0.000	0.005	0.742	0.741	0.694	0.678	0.727	0.674	0.000	0.003	0.862	0.847	0.838	0.809	0.899	0.759	0.000	4
DNT - 0.002	0.859	0.869	0.755	0.447	0.135	0.507	0.000	0.004	0.940	0.921	0.948	0.599	0.031	0.782	0.000	0.001	0.826	0.824	0.792	0.456	0.046	0.508	0.000	
CD8TEM3 - 0.001	0.935	0.936	0.926	0.875	0.309	0 471	0.000	0.004	0.822	0.860	0 877	0.883	0.001	0 5 3 9	0.000	0.002	0.932	0 941	0.922	0.893	0 393	0 714	0.000	- 0.4
CD4Naive/activated = 0.001	0.967	0.910	0.693	0.306	0.135	0.000	0.000	0.001	0.866	0.718	0.813	0.225	0.040	0.000	0.000	0.000	0.744	0 774	0.458	0.296	0.069	0.000	0.000	
TatalManagata 0.001	0.007	0.020	0.709	0.972	0.155	0.000	0.000	0.001	0.000	0.796	0.541	0.225	0.396	0.000	0.000	0.000	0.010	0.011	0.750	0.792	0.005	0.000	0.000	
CDRTEM1/activated 0.001	0.955	0.959	0.750	0.075	0.007	0.005	0.000	0.000	0.041	0.700	0.041	0.017	0.000	0.001	0.000	0.001	0.019	0.011	0.750	0.702	0.762	0.000	0.000	
CDSTEMI/activated = 0.001	0.960	0.980	0.951	0.090	0.507		0.000	0.001	0.341	0.575	0.956	0.000	0.920	0.001	0.000	0.001	0.904	0.955	0.937	0.721	0.555	0.755	0.000	
mDC = 0.000	0.957	0.951	0.941	0.801	0.528	0.678	0.000	0.002	0.786	0.507	0.482	0.338	0.750	0.251	0.000	0.002	0.922	0.912	0.913	0.284	0.519	0.958	0.000	
Th1/activated = 0.000	0.896	0.870	0.812	0.768	0.731	0.084	0.000	0.000	0.962	0.972	0.912	0.891	0.752	0.071	0.000	0.001	0.887	0.896	0.840	0.663	0.821	0.113	0.000	
CD8TEMRA/activated - 0.000	0.914	0.811	0.588	0.547	0.289	0.000	0.000	0.000	0.942	0.894	0.840	0.550	0.525	0.000	0.000	0.000	0.677	0.649	0.468	0.357	0.642	0.000	0.000	- 0.2
Plasmablast - 0.000	0.941	0.946	0.890	0.737	0.017	0.676	0.000	0.000	0.930	0.938	0.760	0.754	0.898	0.837	0.000	0.002	0.811	0.752	0.655	0.405	0.004	0.561	0.000	
Treg/activated - 0.000	0.837	0.808	0.806	0.694	0.000	0.000	0.000	0.000	0.757	0.599	0.608	0.498	0.000	0.000	0.000	0.000	0.634	0.631	0.683	0.308	0.000	0.000	0.000	
CD8Naive/activated - 0.000	0.923	0.892	0.636	0.623	0.067	0.000	0.000	0.000	0.906	0.848	0.807	0.618	0.000	0.000	0.000	0.000	0.584	0.494	0.398	0.217	0.026	0.000	0.000	
CD8TEM2 - 0.000	0.817	0.818	0.769	0.675	0.053	0.034	0.000	0.001	0.878	0.896	0.853	0.702	0.519	0.057	0.000	0.000	0.899	0.883	0.837	0.725	0.078	0.036	0.000	
nnCD4CXCR5pos/activated - 0.000	0.925	0.916	0.757	0.640	0.147	0.000	0.000	0.000	0.824	0.920	0.837	0.722	0.098	0.000	0.000	0.000	0.758			0.496	0.038	0.000	0.000	
pDC - 0.000	0.987	0.983	0.963	0.831	0.980	0.962	0.000	0.001	0.139	0.135	0.133	0.108	0.838	0.157	0.000	0.001	0.926	0.927	0.918	0.590	0.964	0.766	0.000	
DNT/activated - 0.000	0.858	0.827	0.700	0.677	0.179	0.003	0.000	0.000	0.883	0.876	0.840	0.746	0.235	0.000	0.000	0.000	0.563	0.560	0.482	0.415	0.129	0.000	0.000	
CD8TCM/activated = 0.000	0.965	0.953	0.900	0.754	0.000	0.000	0.000	0.000	0.484	0.500	0.497	0.375	0.000	0.000	0.000	0.000	0.964	0.951	0.923	0.599	0.000	0.000	0.000	- 0.0
DPT/activated - 0.000	0.914	0.837	0.569	0.521	0.715	0.000	0.000	0.000	0.250	0.208	0.146	0.150	0.115	0.000	0.000	0.000	0.396	0.407	0.058	0.212	0.041	0.000	0.000	
Th17/activated - 0.000	0.936	0.808	0.741	0.701	0.618	0.001	0.000	0.000	0.941	0.905	0.902	0.755	0.759	0.001	0.000	0.000	0.858	0.793	0.733	0.606	0.751	0.000	0.000	
ILC - 0.000	0.920	0.924	0.872	0.787	0.217	0.002	0.000	0.000	0.534	0.448	0.418	0.334	0.392	0.001	0.000	0.000	0.839	0.806	0.653	0.325	0.228	0.000	0.000	
Th2/activated = 0.000	0.930	0.815	0.535	0.454	0.002	0.000	0.000	0.000	0.943	0.944	0.718	0.588	0.076	0.000	0.000	0.000	0.931	0.893	0.502	0.431	0.000	0.000	0.000	
CD8TEM3/activated = 0.000	0.928	0.919	0.843	0.599	0.002	0.000	0.000	0.000	0.816	0.847	0.782	0.561	0.002	0.000	0.000	0.000	0.942	0.935	0.862	0.540	0.006	0.000	0.000	
CD8TEM2/activated = 0.000	0.672	0.589	0.325	0.089	0.000	0.000	0.000	0.000		0.838	0.567	0.056	0.009	0.000	0.000	0.000	0.944	0.907	0.500	0.130	0.000	0.000	0.000	

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 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 3 (pDC)

ROI 4 (Eosinophil/CD45hiCD66bpos)



ROI 5 (Neutrophil)



ROI 2 (Neutrophil)

ROI 6 (Plasmablast/ CD66bnegCD45lo/IgDnegMemB/EarlyNK)







ROI 7 (CD66bnegCD45lo/ Plasmablast/lgDnegMemB)



ROI 8 (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	Test set		
			(AUROC Mean ± Std.)	(AUROC Mean ± Std.)		
Ν	anual acting	GBDT	0.970 ± 0.091	0.975 ± 0.042		
101	lanual gating	Logistic regression with L2 reg.	0.917 ± 0.047	0.938 ± 0.059		
	FlowSOM	GBDT	0.930 ± 0.089	0.936 ± 0.096		
FlowSOM		Logistic regression with L2 reg.	0.875 ± 0.077	0.902 ± 0.090		
CNN		CNN	0.633 ± 0.274	0.543 ± 0.256		
	Global moon pooling	GBDT	0.890 ± 0.135	0.888 ± 0.127		
		Logistic regression with L2 reg.	0.923 ± 0.064	0.919 ± 0.077		
	Clabel sum neeling	GBDT	0.909 ± 0.126	0.884 ± 0.102		
	Global sum pooling	Logistic regression with L2 reg.	0.910 ± 0.098	0.859 ± 0.155		
Clabel mey realing	GBDT	0.993 ± 0.025	0.963 ± 0.068			
	Global max pooling	Logistic regression with L2 reg.	0.993 ± 0.018	0.989 ± 0.033		
	Clobal min naalina	GBDT	0.942 ± 0.108	0.902 ± 0.127		
	Giobai min pooling	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	Test set		
			(AUROC Mean ± Std.)	(AUROC Mean ± Std.)		
м	anual acting	GBDT	0.735 ± 0.129	0.641 ± 0.154		
IVI	anual gating	Logistic regression with L2 reg.	0.456 ± 0.171	0.446 ± 0.140		
	ElowSOM	GBDT	0.622 ± 0.132	0.579 ± 0.151		
FlowSOM		Logistic regression with L2 reg.	0.577 ± 0.162	0.520 ± 0.167		
	CNN	CNN	0.585 ± 0.157	0.520 ± 0.163		
	Global maan nooling	GBDT	0.605 ± 0.166	0.535 ± 0.136		
	Giobai mean pooling	Logistic regression with L2 reg.	0.626 ± 0.134	0.588 ± 0.125		
	Clobal sum pooling	GBDT	0.613 ± 0.169	0.560 ± 0.132		
avMAE	Giobal sulli poolilig	Logistic regression with L2 reg.	0.643 ± 0.126	0.607 ± 0.138		
CYMAE	Global may pooling	GBDT	0.635 ± 0.146	0.584 ± 0.153		
	Giobai max pooning	Logistic regression with L2 reg.	0.608 ± 0.119	0.625 ± 0.154		
	Global min pooling	GBDT	$0.616\pm0.16\overline{7}$	0.559 ± 0.179		
	Giobai min pooling	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	Test set		
			(AUROC Mean ± Std.)	(AUROC Mean ± Std.)		
М	anual acting	GBDT	0.865 ± 0.126	0.796 ± 0.124		
1 V1	anual gating	Logistic regression with L2 reg.	0.621 ± 0.120	0.615 ± 0.177		
	ElowSOM	GBDT	0.885 ± 0.106	0.859 ± 0.112		
	FIOWSOM	Logistic regression with L2 reg.	0.592 ± 0.134	0.579 ± 0.170		
	CNN	CNN	0.591 ± 0.173	0.531 ± 0.201		
	Global maan nooling	GBDT	0.692 ± 0.171	0.610 ± 0.178		
	Giobai mean pooling	Logistic regression with L2 reg.	0.740 ± 0.158	0.714 ± 0.155		
CYMAE	Clobal sum pooling	GBDT	0.663 ± 0.189	0.651 ± 0.149		
	Giobal sum pooling	Logistic regression with L2 reg.	0.690 ± 0.180	0.676 ± 0.155		

Global may pooling	GBDT	0.750 ± 0.160	0.718 ± 0.152
Giobai max pooning	Logistic regression with L2 reg.	$\textbf{0.895} \pm \textbf{0.103}$	0.887 ± 0.114
Clobal min nealing	GBDT	0.818 ± 0.139	0.821 ± 0.138
Giobal min pooling	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.