Supplementary Materials: SigPrimedNet: a Signaling-informed Neural Network for scRNA-seq Annotation of Known and Unknown Cell Types

Pelin Gundogdu[1,2], Inmaculada Alamo[1,2], Isabel A. Nepomuceno-Chamorro[3], Joaquin Dopazo[1,2,4,5]* and Carlos Loucera[1,2]*

Encoding visualization

The model encodings, i.e., the features learned by the ANN, can be used in a fully unsupervised modality. Thus, given a collection of cells, we can compute their internal representation, and use the features for clustering purposes. To show the representation power of the model we have randomly split the immune dataset using 50% of the samples for training the model and the remaining half for computing the encodings. Figure S1 shows a 2D TSNE visualization of the learned representation for the test.

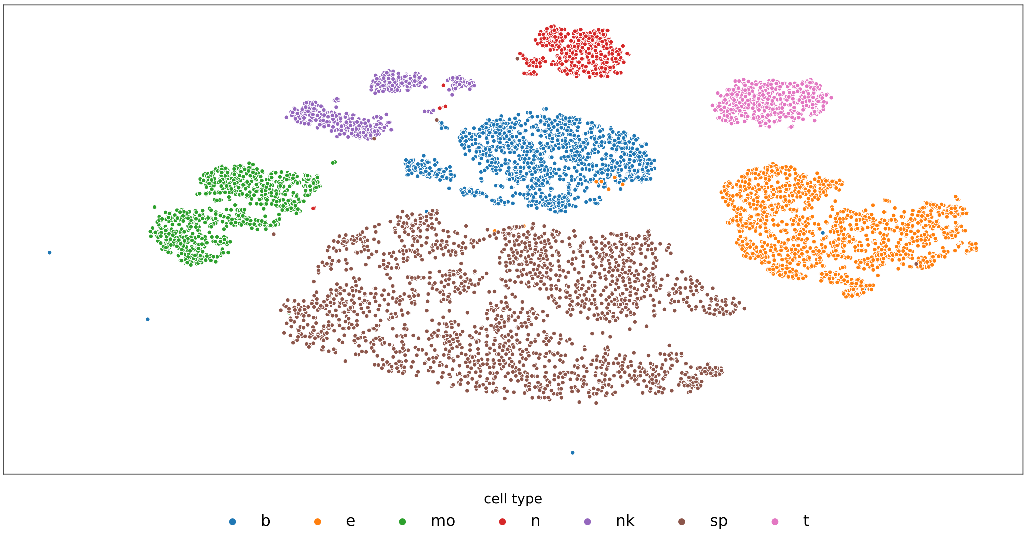
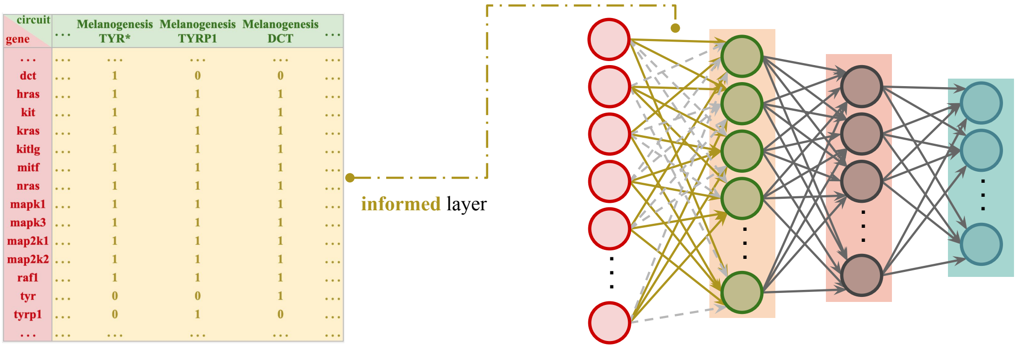


Figure S1. 2D TSNE visualization of the features learned by SigPrimedNet for a test split of the Immune dataset. The cell types b, e, mo, n, nk, sp, and t refer to B cells, erythrocytes, monocytes, neutrophils, NK cells, CD34+ HSPCs, and T cells, respectively.

Two-layer design

In this section, we provide several figures that supplement the information of the main manuscript with respect to SigPrimedNet performance when using a two-layer design, which is built by adding a second non-informed dense hidden layer. See Figure S2 for a visual representation of the two-layer design.



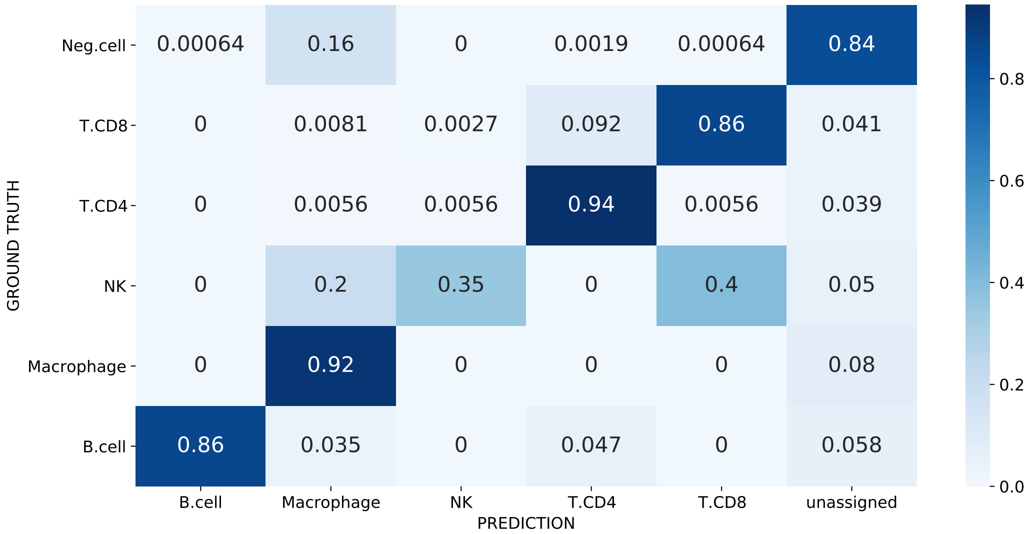
gene \ circuit	…	Melanogenesis TYR*	Melanogenesis TYRP1	Melanogenesis DCT	…
…	…	…	…	…	…
dct	…	1	0	0	…
hras	…	1	1	1	…
kit	…	1	1	1	…
kras	…	1	1	1	…
kitlg	…	1	1	1	…
mitf	…	1	1	1	…
nras	…	1	1	1	…
mapk1	…	1	1	1	…
mapk3	…	1	1	1	…
map2k1	…	1	1	1	…
map2k2	…	1	1	1	…
raf1	…	1	1	1	…
tyr	…	0	0	1	…
tyrp1	…	0	1	0	…
…	…	…	…	…	…

Figure S2. SigPrimedNet with two-layer design.

Interestingly, the most interpretable and less complex model (the one-layer design) is clearly superior to the two-layer design as can be evidenced by observing Table S1 and the confusion matrix S3 (directly comparable to one shown in the main manuscript).

Table S1. Cell type, number of samples detail, and percentage of samples above or below the encoding-based threshold of `Melanoma` dataset during the testing phase. Note that Neg. cells including malignant cells, CAF cells, and endothelial cells were removed from the training set (see Materials).

		1 layer design		*2 layer design*	
Threshold	**Cell type**	**Number of samples**	**Percentage** (%)	**Number of samples**	**Percentage (%)**
above	B.cell	229	93.47	232	94.69
	Macrophage	119	94.44	114	90.48
	NK	27	96.43	26	92.86
	Neg.cell	179	8.03	364	16.34
	T.CD4	243	94.55	247	96.11
	T.CD8	512	96.97	508	96.21
below	B.cell	16	6.53	13	5.31
	Macrophage	7	5.56	12	9.52
	NK	1	3.57	2	7.14
	Neg.cell	2049	91.97	1864	83.66
	T.CD4	14	5.45	10	3.89
	T.CD8	16	3.03	20	3.79



GROUND TRUTH \ PREDICTION	B.cell	Macrophage	NK	T.CD4	T.CD8	unassigned
Neg.cell	0.00064	0.16	0	0.0019	0.00064	0.84
T.CD8	0	0.0081	0.0027	0.092	0.86	0.041
T.CD4	0	0.0056	0.0056	0.94	0.0056	0.039
NK	0	0.2	0.35	0	0.4	0.05
Macrophage	0	0.92	0	0	0	0.08
B.cell	0.86	0.035	0	0.047	0	0.058

Figure S3. The confusion matrix of the `Melanoma` dataset for the unknown cell-type identification task.

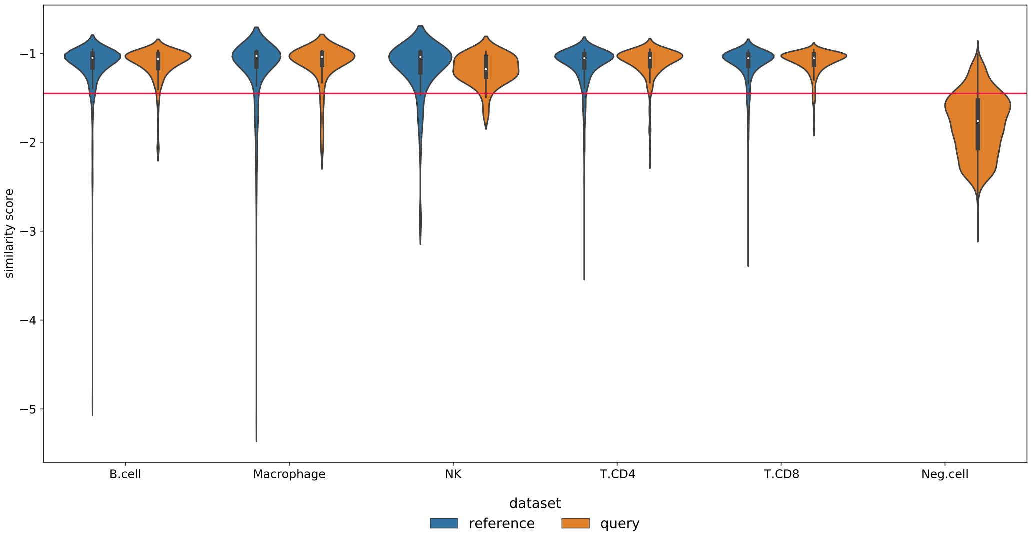
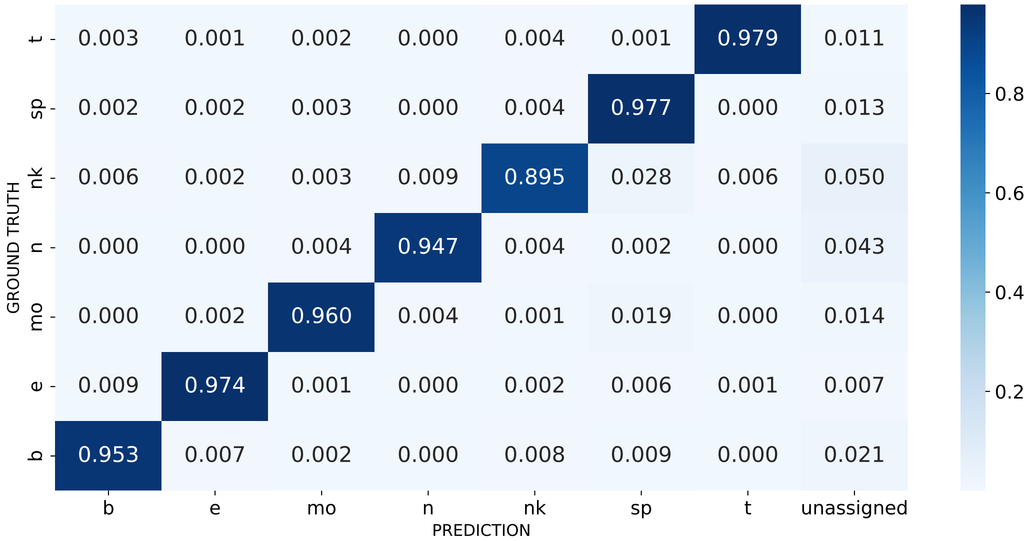


Figure S4. Similarity score distribution for each cell type on the validation and test splits using the two-layer architecture (`Melanoma` dataset). The horizontal line shows the threshold obtained using the reference set inner splits as detailed in the Methods section of the main manuscript.

Here we provide the results of the experiments carried out to test the performance of the SigPrimedNet model using the one-layer design in the main manuscript. The one-layer design has a similar performance to the two-layer design when dealing with tasks where all the cell types are known (see Figures S5, S6 and S7).



	b	e	mo	n	nk	sp	t	unassigned
t	0.003	0.001	0.002	0.000	0.004	0.001	0.979	0.011
sp	0.002	0.002	0.003	0.000	0.004	0.977	0.000	0.013
nk	0.006	0.002	0.003	0.009	0.895	0.028	0.006	0.050
n	0.000	0.000	0.004	0.947	0.004	0.002	0.000	0.043
mo	0.000	0.002	0.960	0.004	0.001	0.019	0.000	0.014
e	0.009	0.974	0.001	0.000	0.002	0.006	0.001	0.007
b	0.953	0.007	0.002	0.000	0.008	0.009	0.000	0.021

Figure S5. The confusion matrix of the `Immune` dataset using SigPrimedNet with 2 layers.

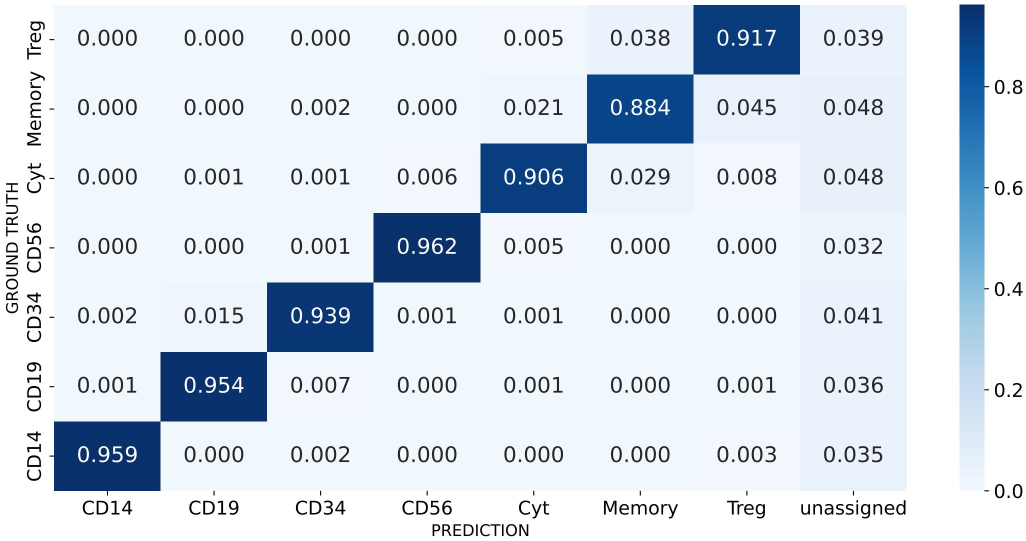
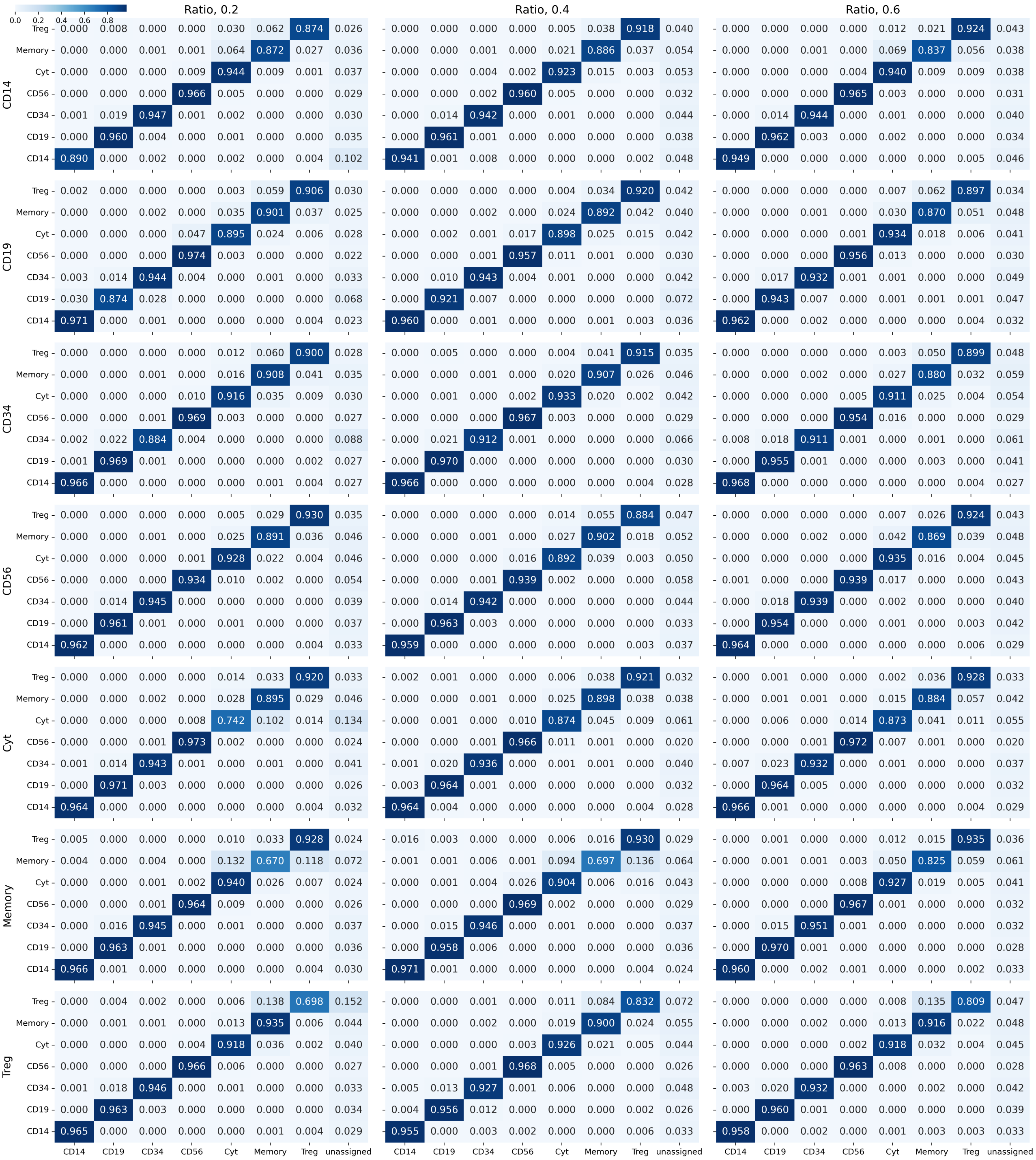


Figure S6. The confusion matrix of the `PBMC` balanced dataset using SigPrimedNet with 2 layers.



Ratio, 0.2

		CD14	CD19	CD34	CD56	Cyt	Memory	Treg	unassigned
CD14	Treg	0.000	0.008	0.000	0.000	0.030	0.062	0.874	0.026
	Memory	0.000	0.000	0.001	0.001	0.064	0.872	0.027	0.036
	Cyt	0.000	0.000	0.000	0.009	0.944	0.009	0.001	0.037
	CD56	0.000	0.000	0.000	0.966	0.005	0.000	0.000	0.029
	CD34	0.001	0.019	0.947	0.001	0.002	0.000	0.000	0.030
	CD19	0.000	0.960	0.004	0.000	0.001	0.000	0.000	0.035
	CD14	0.890	0.000	0.002	0.000	0.002	0.000	0.004	0.102
CD19	Treg	0.002	0.000	0.000	0.000	0.003	0.059	0.906	0.030
	Memory	0.000	0.000	0.002	0.000	0.035	0.901	0.037	0.025
	Cyt	0.000	0.000	0.000	0.047	0.895	0.024	0.006	0.028
	CD56	0.000	0.000	0.000	0.974	0.003	0.000	0.000	0.022
	CD34	0.003	0.014	0.944	0.004	0.000	0.001	0.000	0.033
	CD19	0.030	0.874	0.028	0.000	0.000	0.000	0.000	0.068
	CD14	0.971	0.000	0.001	0.000	0.000	0.000	0.004	0.023
CD34	Treg	0.000	0.000	0.000	0.000	0.012	0.060	0.900	0.028
	Memory	0.000	0.000	0.001	0.000	0.016	0.908	0.041	0.035
	Cyt	0.000	0.000	0.000	0.010	0.916	0.035	0.009	0.030
	CD56	0.000	0.000	0.001	0.969	0.003	0.000	0.000	0.027
	CD34	0.002	0.022	0.884	0.004	0.000	0.000	0.000	0.088
	CD19	0.001	0.969	0.001	0.000	0.000	0.000	0.002	0.027
	CD14	0.966	0.000	0.000	0.000	0.000	0.001	0.004	0.027
CD56	Treg	0.000	0.000	0.000	0.000	0.005	0.029	0.930	0.035
	Memory	0.000	0.000	0.001	0.000	0.025	0.891	0.036	0.046
	Cyt	0.000	0.000	0.000	0.001	0.928	0.022	0.004	0.046
	CD56	0.000	0.000	0.000	0.934	0.010	0.002	0.000	0.054
	CD34	0.000	0.014	0.945	0.000	0.000	0.000	0.000	0.039
	CD19	0.000	0.961	0.001	0.000	0.001	0.000	0.000	0.037
	CD14	0.962	0.000	0.000	0.000	0.000	0.000	0.004	0.033
Cyt	Treg	0.000	0.000	0.000	0.000	0.014	0.033	0.920	0.033
	Memory	0.000	0.000	0.002	0.000	0.028	0.895	0.029	0.046
	Cyt	0.000	0.000	0.000	0.008	0.742	0.102	0.014	0.134
	CD56	0.000	0.000	0.001	0.973	0.002	0.000	0.000	0.024
	CD34	0.001	0.014	0.943	0.001	0.000	0.001	0.000	0.041
	CD19	0.000	0.971	0.003	0.000	0.000	0.000	0.000	0.026
	CD14	0.964	0.000	0.000	0.000	0.000	0.000	0.004	0.032
Memory	Treg	0.005	0.000	0.000	0.000	0.010	0.033	0.928	0.024
	Memory	0.004	0.000	0.004	0.000	0.132	0.670	0.118	0.072
	Cyt	0.000	0.000	0.001	0.002	0.940	0.026	0.007	0.024
	CD56	0.000	0.000	0.001	0.964	0.009	0.000	0.000	0.026
	CD34	0.000	0.016	0.945	0.000	0.001	0.000	0.000	0.037
	CD19	0.000	0.963	0.001	0.000	0.000	0.000	0.000	0.036
	CD14	0.966	0.001	0.000	0.000	0.000	0.000	0.004	0.030
Treg	Treg	0.000	0.004	0.002	0.000	0.006	0.138	0.698	0.152
	Memory	0.000	0.001	0.001	0.000	0.013	0.935	0.006	0.044
	Cyt	0.000	0.000	0.000	0.004	0.918	0.036	0.002	0.040
	CD56	0.000	0.000	0.000	0.966	0.006	0.000	0.000	0.027
	CD34	0.001	0.018	0.946	0.000	0.001	0.000	0.000	0.033
	CD19	0.000	0.963	0.003	0.000	0.000	0.000	0.000	0.034
	CD14	0.965	0.000	0.000	0.000	0.000	0.001	0.004	0.029

Ratio, 0.4

		CD14	CD19	CD34	CD56	Cyt	Memory	Treg	unassigned
CD14	Treg	0.000	0.000	0.000	0.000	0.005	0.038	0.918	0.040
	Memory	0.000	0.000	0.001	0.000	0.021	0.886	0.037	0.054
	Cyt	0.000	0.000	0.004	0.002	0.923	0.015	0.003	0.053
	CD56	0.000	0.000	0.002	0.960	0.005	0.000	0.000	0.032
	CD34	0.000	0.014	0.942	0.000	0.001	0.000	0.000	0.044
	CD19	0.000	0.961	0.001	0.000	0.000	0.000	0.000	0.038
	CD14	0.941	0.001	0.008	0.000	0.000	0.000	0.002	0.048
CD19	Treg	0.000	0.000	0.000	0.000	0.004	0.034	0.920	0.042
	Memory	0.000	0.000	0.002	0.000	0.024	0.892	0.042	0.040
	Cyt	0.000	0.002	0.001	0.017	0.898	0.025	0.015	0.042
	CD56	0.000	0.000	0.001	0.957	0.011	0.001	0.000	0.030
	CD34	0.000	0.010	0.943	0.004	0.001	0.000	0.000	0.042
	CD19	0.000	0.921	0.007	0.000	0.000	0.000	0.000	0.072
	CD14	0.960	0.000	0.001	0.000	0.000	0.001	0.003	0.036
CD34	Treg	0.000	0.005	0.000	0.000	0.004	0.041	0.915	0.035
	Memory	0.000	0.000	0.001	0.000	0.020	0.907	0.026	0.046
	Cyt	0.000	0.001	0.000	0.002	0.933	0.020	0.002	0.042
	CD56	0.000	0.000	0.000	0.967	0.003	0.000	0.000	0.029
	CD34	0.000	0.021	0.912	0.001	0.000	0.000	0.000	0.066
	CD19	0.000	0.970	0.000	0.000	0.000	0.000	0.000	0.030
	CD14	0.966	0.000	0.000	0.000	0.000	0.000	0.004	0.028
CD56	Treg	0.000	0.000	0.000	0.000	0.014	0.055	0.884	0.047
	Memory	0.000	0.000	0.001	0.000	0.027	0.902	0.018	0.052
	Cyt	0.000	0.000	0.000	0.016	0.892	0.039	0.003	0.050
	CD56	0.000	0.000	0.001	0.939	0.002	0.000	0.000	0.058
	CD34	0.000	0.014	0.942	0.000	0.000	0.000	0.000	0.044
	CD19	0.000	0.963	0.003	0.000	0.000	0.000	0.000	0.033
	CD14	0.959	0.000	0.000	0.000	0.000	0.000	0.003	0.037
Cyt	Treg	0.002	0.001	0.000	0.000	0.006	0.038	0.921	0.032
	Memory	0.000	0.000	0.001	0.000	0.025	0.898	0.038	0.038
	Cyt	0.000	0.001	0.000	0.010	0.874	0.045	0.009	0.061
	CD56	0.000	0.000	0.001	0.966	0.011	0.001	0.000	0.020
	CD34	0.001	0.020	0.936	0.000	0.001	0.001	0.000	0.040
	CD19	0.003	0.964	0.001	0.000	0.000	0.000	0.000	0.032
	CD14	0.964	0.004	0.000	0.000	0.000	0.000	0.004	0.028
Memory	Treg	0.016	0.003	0.000	0.000	0.006	0.016	0.930	0.029
	Memory	0.001	0.001	0.006	0.001	0.094	0.697	0.136	0.064
	Cyt	0.000	0.001	0.004	0.026	0.904	0.006	0.016	0.043
	CD56	0.000	0.000	0.000	0.969	0.002	0.000	0.000	0.029
	CD34	0.000	0.015	0.946	0.000	0.001	0.000	0.000	0.037
	CD19	0.000	0.958	0.006	0.000	0.000	0.000	0.000	0.036
	CD14	0.971	0.001	0.000	0.000	0.000	0.000	0.004	0.024
Treg	Treg	0.000	0.000	0.001	0.000	0.011	0.084	0.832	0.072
	Memory	0.000	0.000	0.002	0.000	0.019	0.900	0.024	0.055
	Cyt	0.000	0.000	0.000	0.003	0.926	0.021	0.005	0.044
	CD56	0.000	0.000	0.001	0.968	0.005	0.000	0.000	0.026
	CD34	0.005	0.013	0.927	0.001	0.006	0.000	0.000	0.048
	CD19	0.004	0.956	0.012	0.000	0.000	0.000	0.002	0.026
	CD14	0.955	0.000	0.003	0.002	0.000	0.000	0.006	0.033

Ratio, 0.6

		CD14	CD19	CD34	CD56	Cyt	Memory	Treg	unassigned
CD14	Treg	0.000	0.000	0.000	0.000	0.012	0.021	0.924	0.043
	Memory	0.000	0.000	0.001	0.000	0.069	0.837	0.056	0.038
	Cyt	0.000	0.000	0.000	0.004	0.940	0.009	0.009	0.038
	CD56	0.000	0.000	0.000	0.965	0.003	0.000	0.000	0.031
	CD34	0.000	0.014	0.944	0.000	0.001	0.000	0.000	0.040
	CD19	0.000	0.962	0.003	0.000	0.002	0.000	0.000	0.034
	CD14	0.949	0.000	0.000	0.000	0.000	0.000	0.005	0.046
CD19	Treg	0.000	0.000	0.000	0.000	0.007	0.062	0.897	0.034
	Memory	0.000	0.000	0.001	0.000	0.030	0.870	0.051	0.048
	Cyt	0.000	0.000	0.000	0.001	0.934	0.018	0.006	0.041
	CD56	0.000	0.000	0.000	0.956	0.013	0.000	0.000	0.030
	CD34	0.000	0.017	0.932	0.001	0.001	0.000	0.000	0.049
	CD19	0.000	0.943	0.007	0.000	0.001	0.001	0.001	0.047
	CD14	0.962	0.000	0.002	0.000	0.000	0.000	0.004	0.032
CD34	Treg	0.000	0.000	0.000	0.000	0.003	0.050	0.899	0.048
	Memory	0.000	0.000	0.002	0.000	0.027	0.880	0.032	0.059
	Cyt	0.000	0.000	0.000	0.005	0.911	0.025	0.004	0.054
	CD56	0.000	0.000	0.000	0.954	0.016	0.000	0.000	0.029
	CD34	0.008	0.018	0.911	0.001	0.000	0.001	0.000	0.061
	CD19	0.000	0.955	0.001	0.000	0.000	0.003	0.000	0.041
	CD14	0.968	0.000	0.000	0.000	0.000	0.000	0.004	0.027
CD56	Treg	0.000	0.000	0.000	0.000	0.007	0.026	0.924	0.043
	Memory	0.000	0.000	0.002	0.000	0.042	0.869	0.039	0.048
	Cyt	0.000	0.000	0.000	0.000	0.935	0.016	0.004	0.045
	CD56	0.001	0.000	0.000	0.939	0.017	0.000	0.000	0.043
	CD34	0.000	0.018	0.939	0.000	0.002	0.000	0.000	0.040
	CD19	0.000	0.954	0.000	0.000	0.001	0.000	0.003	0.042
	CD14	0.964	0.000	0.000	0.000	0.000	0.000	0.006	0.029
Cyt	Treg	0.000	0.001	0.000	0.000	0.002	0.036	0.928	0.033
	Memory	0.000	0.001	0.001	0.000	0.015	0.884	0.057	0.042
	Cyt	0.000	0.006	0.000	0.014	0.873	0.041	0.011	0.055
	CD56	0.000	0.000	0.001	0.972	0.007	0.001	0.000	0.020
	CD34	0.007	0.023	0.932	0.000	0.001	0.000	0.000	0.037
	CD19	0.000	0.964	0.005	0.000	0.000	0.000	0.000	0.032
	CD14	0.966	0.001	0.000	0.000	0.000	0.000	0.004	0.029
Memory	Treg	0.000	0.001	0.000	0.000	0.012	0.015	0.935	0.036
	Memory	0.000	0.001	0.001	0.003	0.050	0.825	0.059	0.061
	Cyt	0.000	0.000	0.000	0.008	0.927	0.019	0.005	0.041
	CD56	0.000	0.000	0.000	0.967	0.001	0.000	0.000	0.032
	CD34	0.000	0.015	0.951	0.001	0.000	0.000	0.000	0.032
	CD19	0.000	0.970	0.001	0.000	0.000	0.000	0.000	0.028
	CD14	0.960	0.000	0.002	0.001	0.000	0.000	0.002	0.033
Treg	Treg	0.000	0.000	0.000	0.000	0.008	0.135	0.809	0.047
	Memory	0.000	0.000	0.002	0.000	0.013	0.916	0.022	0.048
	Cyt	0.000	0.000	0.000	0.002	0.918	0.032	0.004	0.045
	CD56	0.000	0.000	0.000	0.963	0.008	0.000	0.000	0.028
	CD34	0.003	0.020	0.932	0.000	0.000	0.002	0.000	0.042
	CD19	0.000	0.960	0.001	0.000	0.000	0.000	0.000	0.039
	CD14	0.958	0.000	0.002	0.000	0.001	0.003	0.003	0.033

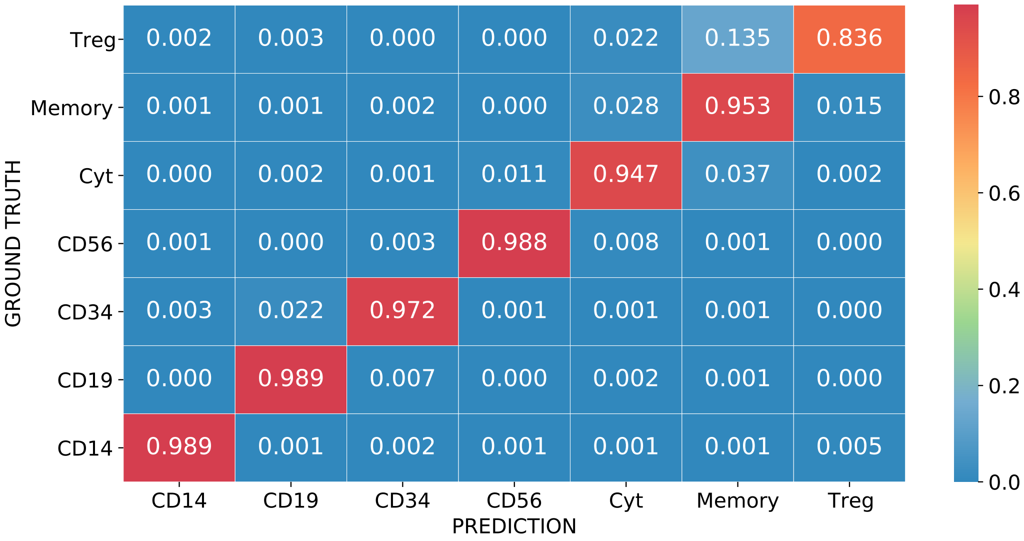
Figure S7. The confusion matrix of the `PBMC` unbalanced dataset using SigPrimedNet with 2 layers.

Supervised performance

In this section we present a reduced version of our model (by only keeping SigPrimedNet supervised capabilities) in order to facilitate the comparison of our model with other fully supervised models, i.e. those that do not have a way to label cells as of unknown type.

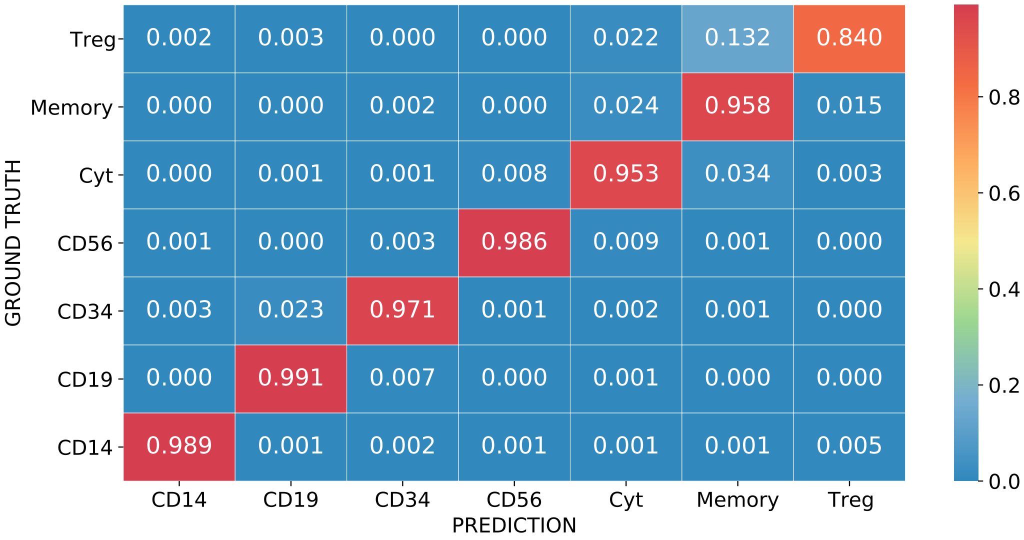
Synthetically balanced PBMC

Following the experiments of the main manuscript, we use a 50 times repeated stratified by cell type 10-fold cross-validation schema using the *balanced* `PBMC` dataset to evaluate the supervised performance of the reduced model. The aggregated confusion matrix exhibits a great ability to correctly assign cell types, as can be seen in Figure S8 (see Figure S9 for 2-layer design). Figure S10 shows the distribution across the test sets of standard classification metrics F1, recall, and precision along with accuracy and balanced accuracy. Whereas, in Figure S11 we can observe in more detail the F1, precision, and recall for each cell type. As expected, the reduced model shows a similar performance to that of the full model when dealing with known cell types (obviously it lacks unknown-cell identification by design), and also shares the same deficiencies when dealing with closely related cell types.



GROUND TRUTH \ PREDICTION	CD14	CD19	CD34	CD56	Cyt	Memory	Treg
Treg	0.002	0.003	0.000	0.000	0.022	0.135	0.836
Memory	0.001	0.001	0.002	0.000	0.028	0.953	0.015
Cyt	0.000	0.002	0.001	0.011	0.947	0.037	0.002
CD56	0.001	0.000	0.003	0.988	0.008	0.001	0.000
CD34	0.003	0.022	0.972	0.001	0.001	0.001	0.000
CD19	0.000	0.989	0.007	0.000	0.002	0.001	0.000
CD14	0.989	0.001	0.002	0.001	0.001	0.001	0.005

Figure S8. PBMC experiment aggregated cross-validation confusion matrix for SigPrimedNet (1-layer design).



GROUND TRUTH \ PREDICTION	CD14	CD19	CD34	CD56	Cyt	Memory	Treg
Treg	0.002	0.003	0.000	0.000	0.022	0.132	0.840
Memory	0.000	0.000	0.002	0.000	0.024	0.958	0.015
Cyt	0.000	0.001	0.001	0.008	0.953	0.034	0.003
CD56	0.001	0.000	0.003	0.986	0.009	0.001	0.000
CD34	0.003	0.023	0.971	0.001	0.002	0.001	0.000
CD19	0.000	0.991	0.007	0.000	0.001	0.000	0.000
CD14	0.989	0.001	0.002	0.001	0.001	0.001	0.005

Figure S9. PBMC experiment aggregated cross-validation confusion matrix for SigPrimedNet (2-layer design).

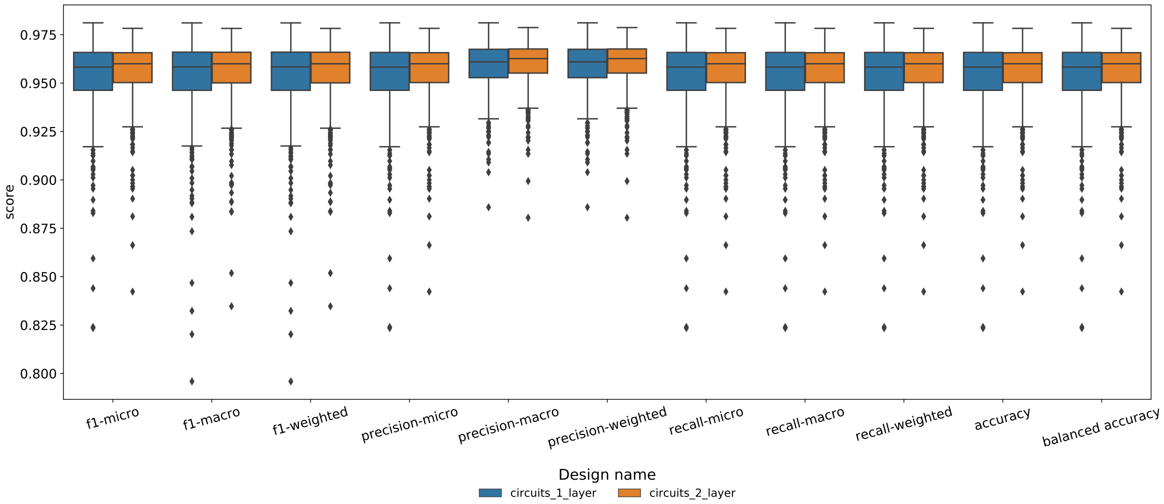


Figure S10. Performance of SigPrimedNet (1 and 2 layer designs) for the `PBMC` experiment: F1, Precision and Recall score distribution across the test sets of 50 times repeated 10-fold cross-validation.

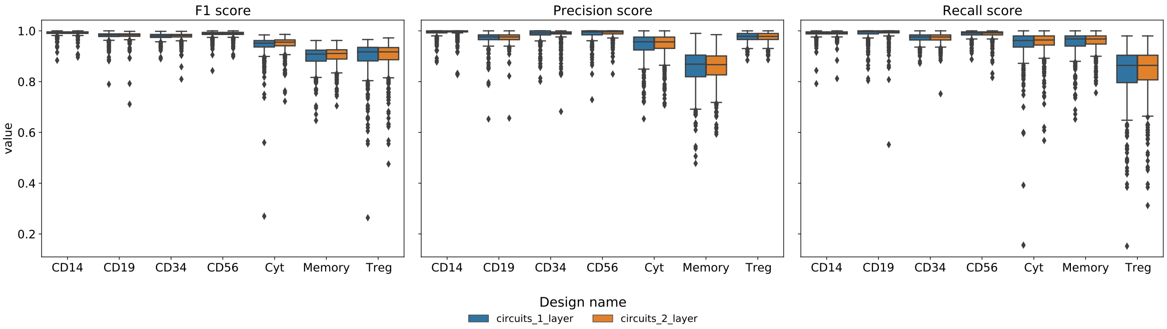
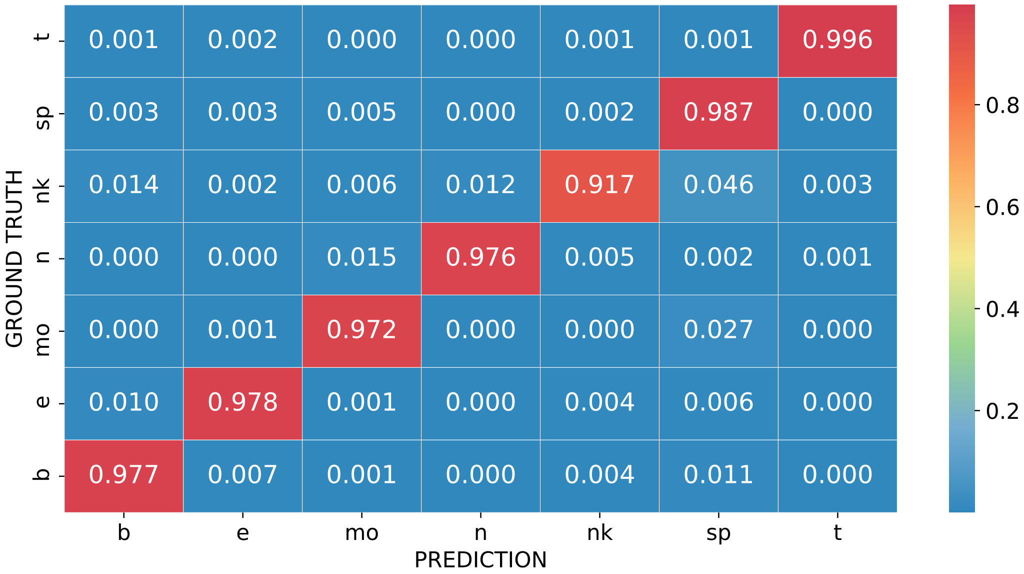


Figure S11. Performance of SigPrimedNet (1 and 2 layer designs) for the `PBMC` experiment: F1, Precision and Recall score distribution across each cell type of the test set of 50 times repeated 10-fold cross-validation.

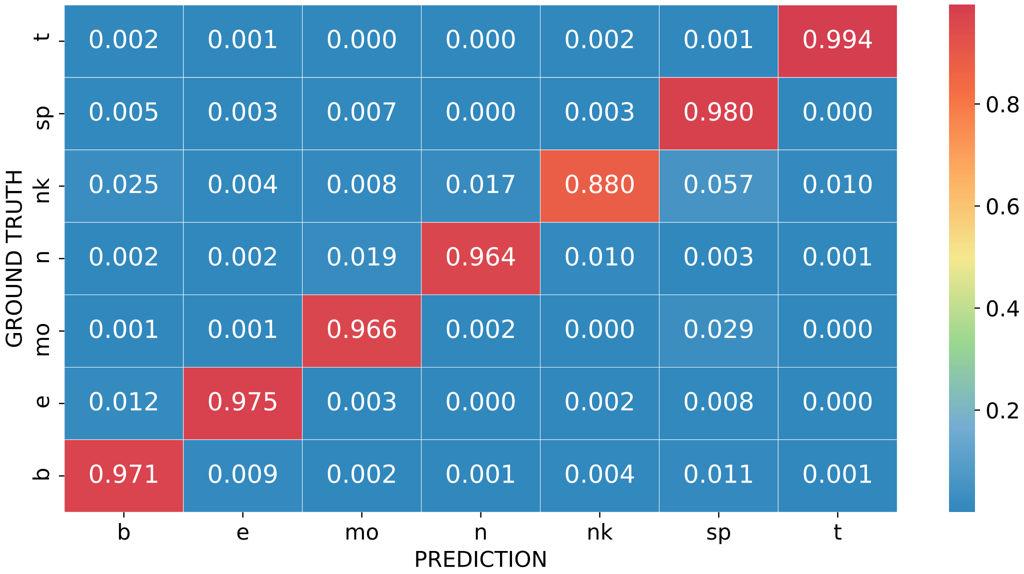
Real-world unbalanced scenario

Following the experiments described in the main manuscript we have used 30 times cell-type stratified repeated 10-fold cross-validation schema using the `Inmune` dataset in order to test the supervised performance of the reduced model. Although the dataset cell type populations are unbalanced, the reduced model can correctly be observed in the aggregated confusion matrix depicted for the reduced model (1-layer design) in Figure S12 (See Figure S13 for the 2-alyer design). Figures S14 and S15 show the general performance of the method (1- and 2-layer designs) and the F1, precision, and recall distribution across the tests sets for each cell type. Most miss-classifications are found in B cells predicted as HSPCs, which could be related to either their shared proliferative capabilities or the overrepresentation of HSPCs in the dataset. Interestingly, the more interpretable 1-layer design outperforms the 2-layer design when dealing with NK cells, which further reinforces the results observed in the main manuscript for the full model.



	b	e	mo	n	nk	sp	t
t	0.001	0.002	0.000	0.000	0.001	0.001	0.996
sp	0.003	0.003	0.005	0.000	0.002	0.987	0.000
nk	0.014	0.002	0.006	0.012	0.917	0.046	0.003
n	0.000	0.000	0.015	0.976	0.005	0.002	0.001
mo	0.000	0.001	0.972	0.000	0.000	0.027	0.000
e	0.010	0.978	0.001	0.000	0.004	0.006	0.000
b	0.977	0.007	0.001	0.000	0.004	0.011	0.000

Figure S12. The aggregated confusion matrix of the `Immune` dataset (1-layer design for the reduced model).



	b	e	mo	n	nk	sp	t
t	0.002	0.001	0.000	0.000	0.002	0.001	0.994
sp	0.005	0.003	0.007	0.000	0.003	0.980	0.000
nk	0.025	0.004	0.008	0.017	0.880	0.057	0.010
n	0.002	0.002	0.019	0.964	0.010	0.003	0.001
mo	0.001	0.001	0.966	0.002	0.000	0.029	0.000
e	0.012	0.975	0.003	0.000	0.002	0.008	0.000
b	0.971	0.009	0.002	0.001	0.004	0.011	0.001

Figure S13. The aggregated confusion matrix of the `Immune` dataset (2-layer design for the reduced model).

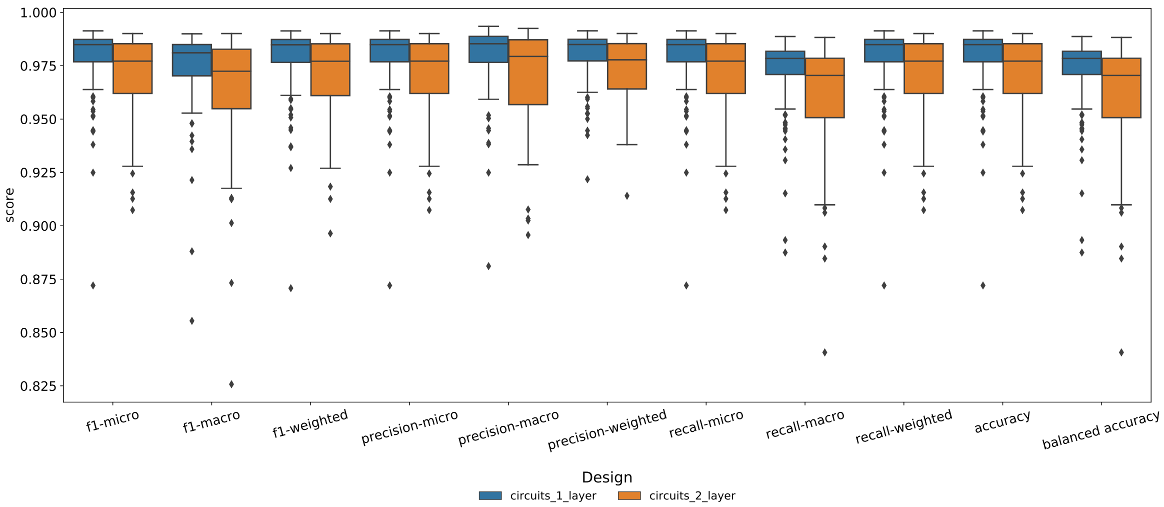


Figure S14. SigPrimedNet overall performance for `Immune` dataset.

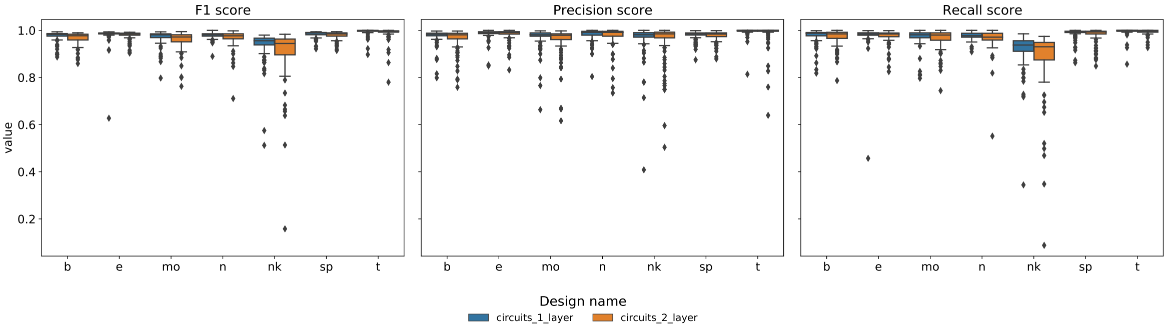


Figure S15. SigPrimedNet performance desegregated for each cell type for `Immune` dataset.