CyAnno: A semi-automated approach for cell type annotation of mass cytometry datasets

Abhinav Kaushik[1], Diane Dunham[1], Ziyuan He[1], Monali Manohar[1], Manisha Desai[2], Kari C Nadeau[1], Sandra Andorf [1,3,4*]

[1] Sean N Parker Center for Allergy and Asthma Research at Stanford University, Stanford University, Stanford, CA 94305–5101, USA
[2] Quantitative Sciences Unit, Stanford University, Stanford, CA 94305–5101, USA
[3] Department of Pediatrics, University of Cincinnati College of Medicine, Cincinnati, OH, USA
[4] Divisions of Biomedical Informatics and Allergy & Immunology, Cincinnati Children's Hospital Medical Center, Cincinnati, OH, USA

***** To whom correspondence should be addressed.

Supplementary information

1 Dataset description

Several datasets were used in this study for profiling the ungated cell, method development and result evaluation, viz. Levine 13 dim (Levine *et al.*, 2015), Levine 32 dim (Levine *et al.*, 2015), Samusik (Samusik *et al.*, 2016), Multi-center datasets (Nassar *et al.*, 2015) and POISED (Table 1). The Samusik and Multi-center datasets were used for both method development and its evaluation, while POISED was used as an independent dataset for the method's evaluation. The Levine 13 dim is a CyTOF dataset in which samples were taken from human bone marrow cells of a single healthy donor, with a panel composed of 13 protein markers and 24 manually gated cell types. Similarly, Levine 32 dim is a CyTOF dataset composed of 32 protein markers with 14 manually gated cell types from human bone marrow cells from two healthy donors (Levine *et al.*, 2015). The "Samusik" (or PANORAMA) CyTOF dataset was obtained from (Samusik *et al.*, 2016) and contains samples from bone marrow of 10 different mice in which the expression profile is measured using a panel of 39 markers. The "Multi-center" dataset contains 16 samples and a panel of 26 markers, wherein only 8 were used for building the classification models, and samples were run across two different centers (or batches) (Nassar *et al.*, 2015). In this work, we utilized the same set of lineage markers for the Samusik and Multi-center datasets as used in the original publications. While the Samusik dataset contains data of 24 different cell populations, the Multi-center dataset contains only four cell types. In the Samusik dataset, the cell type HSC had a very small cell count (< 10 cells; mean ~3 cells per sample) in the overall dataset and was therefore excluded from the training set, however, the HSC cells remained as part of the test set. The publicly available CyTOF datasets used in this study were obtained from the respective publications or HDCytoData R package (Weber and Soneson, 2019). Despite being commonly used in many studies, we found that these and other public datasets with

manually gated cell type labels have one or more of the following limitations for testing any cell type prediction algorithm. The shortcomings include small sample size, small number of batches (≤ 2), mostly distinct with only few closely related cell types, and no varying biological treatments. In fact, the limited sample size did not allow us to perform exhaustive model evaluation on large independent datasets belonging to untrained batches or stimulations. Therefore, we relied on an additional dataset (i.e. POISED (Chinthrajah *et al.*, 2019)) of 30 CyTOF samples from human PBMCs, from peanut-allergic individuals, for testing our algorithm. These samples were run across 7 batches with a panel of 21 lineage markers and 18 functional markers (Table S1) and under two different stimulations, i.e., unstimulated and peanut stimulated. 21 cell types were manually gated (including closely related cell types). The non-canonical peanut reactive T cells used in this study were CD69+ CD40L+ CD4+ T cells and CD69+ CD8+ T cells (Neeland *et al.*, 2020).

2 POISED dataset: Experimental details

2.1 Sample collection

From each study participant, 40 mL blood was drawn by venipuncture at baseline, and at various time points over the course of peanut OIT. Using validated, standardized, and published procedures (Fuss *et al.*, 2009), PBMCs and plasma isolated by ficoll-based density gradient centrifugation were frozen in aliquots and stored in liquid nitrogen and at -80°C, respectively, until thawing for respective assays. In this analysis, we used samples taken at baseline (i.e., during the initial screening phase or at week 0 post-enrollment).

2.2 Mass cytometry

The detailed panel of metal-conjugated antibodies used for mass cytometry analysis of PBMCs is shown in Supplemental Table S1. In-house metal conjugation of purified antibodies as indicated in this table was carried out using Maxpar antibody labeling kits (Fluidigm Co., South San Francisco, CA). PBMCs were thawed and rested overnight at 37°C with 5% CO2 in a combination of RPMI, 10% FBS, and Pen-Strep. Cells were plated in a round-bottom, 96-well plate at the density of 3x106 cells in culture with 300 μL RPMI, 5% FBS, Pen-Strep per well. To evaluate peanut-specific response, PBMCs were stimulated with 200 μg/mL peanut solution for 24 hours with the addition of brefeldin A (5 μg/mL; Biolegend, San Diego, CA) for the last 4 hours. Unstimulated cells treated with brefeldin A for 4 hours before harvesting served as controls. Cells were harvested and stained as previously published (Lin *et al.*, 2015) with an added barcoding step post-permeabilizing the cells. Cell-ID 20-plex Palladium barcoding set (Fluidigm Co.) was used for barcoding according to manufacturer's instructions. Barcoded and pooled samples were acquired on a Helios mass cytometer (Fluidigm Co.). Data normalization, concatenation, and debarcoding were performed using CyTOF software (Fluidigm Co.). Manual gating on raw data files was carried out using FlowJo v10 (FlowJo LLC., Ashland, OR).

3 Data availability

The publicly available datasets can be obtained from HDCytoData R package (Weber and Soneson, 2019). Multi-center dataset can be obtained as publicly available datasets from GitHub (Abdelaal *et al.*, 2019) (https://github.com/tabdelaal/CyTOF-Linear-Classifier). Raw unlabeled POISED dataset files in FCS 3.0 format and normalized labelled CSV format can be obtained from FlowRepository (FR-FCM-Z2V9) (Spidlen *et al.*, 2012).

4 ML parameter tuning and error estimation

For XGboost, decision tree algorithm with base learner "*gbtree*" is used with learning task objective set to "*binary:logistic*" for binary classification per cell type. For random grid search space, the set of gamma $\in$ {0.5, 1.0, 1.5, 2.0, 5.0, 10}, validation subsampling $\in$ {0.3, 0.2} and the learning rate $\in$ {0.01, 0.1, 0.3} are used with maximum tree depth of 6. For feature selection, "*thrifty*" selector with setting of top 5 features per group are used, with L1 regularization weight (*reg_alpha*) set to 0.005. Models are evaluated using validation dataset with two different evaluation matrices- logloss matrix and binary classification error rate (#FP / #(FP+TP)) in which the cell label prediction score larger than 0.6 are used as positive instances. Usually, the prediction score of >0.5 for the binary classification should be sufficient, however, the score of 0.6 ensures for higher precision in outcome. Early stopping parameter is also used, in which model training terminates if validation score does not improve in 10 consecutive iterations. For MLP, Stochastic gradient descent (*sgd*) solver is used for weight optimization with adaptive learning rate. For hyper-parameter optimization, the sets of L2 regularization term $\in$ {1, 0.1, 0.01} and initial learning rate $\in$ {0.01, 0.1, 1.0} are used with validation subsampling $\in$ {0.3, 0.2}. Early stopping parameter is set to 10 which terminates model training if validation score does not improve by value of 0.001 in 10 consecutive iterations. For SVM classifier, Radial Basis Function (*rbf*) kernel trick is used for building non-linear classification boundaries. Here, hyper-parameters set of $C \in \{0.1, 0.01, 0.001\}$ and $\gamma \in \{1, 0.1, 0.01\}$ are used for balancing SVM classification error, regularization and bandwidth estimation. In all of the cases, a total of 10 random hyper-parameters combinations are used for model training and evaluation.

5 Comparison with existing methods

To evaluate the efficacy of CyAnno, we applied the proposed pipeline to the Samusik, Multi-center and POISED CyTOF datasets. The three methods (CyAnno, DeepCyTOF, LDA) were tested across five independent runs, in which, during each run, 20% of the randomly selected samples were used for training the model (i.e., training set), whereas the rest of the samples were evaluated for model testing (i.e., test set). During each run, a F1 score (see below; section 6) is computed for each sample and after five independent runs the mean F1 score for each sample of a given dataset was computed as,

$$\overline{f_s} = \frac{\sum_{i=1}^{i=t} f_i^s}{t}$$

Where, $\overline{f_s}$ is the mean F1 score (*f*) for each sample *s* in a given dataset averaged across *t* (here *t*=5) independent runs. Wherein, TP 'ungated' cells were excluded for estimating the F1 score, in order to avoid any biasness in evaluating the efficacy of different methods in predicting the gated cell populations.

Next, we also computed the mean F1 score for each cell type per dataset using the following equation:

$$\bar{f_c} = \frac{\sum_{i=1}^{i=(t\times ns)} f_i^c}{(t \times ns)}$$

Where, $\bar{f_c}$ is the mean F1 score (*f*) computed for each cell type *c* in a dataset with *ns* samples across *t* independent runs (here *t*=5). In an additional analysis (Figure 3), the F1 scores for each sample or cell type were also computed without ungated live cells in the test sample. For CyAnno, we used ungated cells in the training set for cell type classification, unlike the LDA and DeepCyTOF methods which did not allow the inclusion of ungated cells for model training.

6 Performance matrices

The CyAnno performance was compared and evaluated in terms of F1 score which was defined as:

$$F1 = 2 \times \frac{precision \times recall}{precion + recall}$$

Where, $precision = TP/(TP + FP)$ and $recall = TP/(TP + FN)$. For a given dataset (e.g., Samusik), we estimated the F1 score of each sample. Similarly, we also computed the mean F1 score for each cell type, across different samples of a given dataset (e.g., 20 samples of the POISED dataset).

To further confirm the accuracy of built models in predicting correct cell labels, we also performed a permutation test of predicted labels to calculate the significance of predictions. Wherein, the classification accuracy of CyAnno was compared after randomly permuting the cell label *1000* times and examining the change in sample F1 score. The statistical significance was measured by *p* value which reflected the probability of obtaining a high F1 by chance, and calculated as:

$$p = \frac{\Delta\{F1^{perm} \geq F1\}}{n}$$

Where, Δ is the cardinality of set when the F1 scores of permuted samples were larger than or equal to the F1 score obtained with original sample cell labels. Low *p* value (e.g., $p < 0.05$) suggest high accuracy in correctly predicting cell labels.

7 ML classifiers evaluation

7.1 Computation cost

Since three different CTSMs can be built for each cell type, or they can be used together with the ensemble method, the total computational cost for running CyAnno is directly proportional to the number of cells, cell types and ML algorithms used. Figure S11A depicts the total time taken to build an optimized CTSM based on XGboost, MLP and SVM for each cell type in the POISED dataset. Here most of the time is used in hyper-parameter optimization which is an essential part of any ML based pipeline. For faster results, we found that XGboost outperforms any of the other ML algorithms used. Although, for highest accuracy we recommend using ensemble classifier as it uses prediction results from multiple ML algorithms and thus less likely to produce over-fitted results.

7.2 Evaluation of different ML classifiers

Figure S11B-D shows the F1 score achieved by CyAnno with different classifiers for predicting cells of each cell type in the test datasets, wherein XGboost outperforms both MLP and SVM in predicting the correct cell type labels. However, for the evaluation, instead of the single XGboost classifier, we used the consensus of prediction results from all of the classifiers using a majority voting approach to build a cell type specific ensemble classifier. Such Ensemble or hybrid classifiers are less likely to misclassify unseen data and also less probable to over-fit the training set (Ranawana and Palade, 2016; Hsu, 2017).

Supplementary Table S1. Panel of metal-conjugated antibodies used for mass cytometry analysis of PBMCs for POISED dataset.

Marker	antibody	Marker type
CD19	Nd142Di	Lineage
CD49b	Nd143Di	Lineage
CD4	Nd145Di	Lineage
CD8	Nd146Di	Lineage
CD20	Sm147Di	Lineage
CD38	Nd148Di	Lineage
CCR4	Sm149Di	Lineage
LAG3	Nd150Di	Lineage
CD123	Eu151Di	Lineage
CD45RA	Eu153Di	Lineage
CD3	Sm154Di	Lineage
HLA-DR	Gd157Di	Lineage
CD33	Gd158Di	Lineage
CD11c	Tb159Di	Lineage
CD14	Gd160Di	Lineage
CD69	Dy162Di	Lineage (To identify pea-specific cells)
CXCR3	Dy163Di	Lineage
CD127	Ho165Di	Lineage
CD27	Er167Di	Lineage
CD40L	Er168Di	Lineage (To identify pea-specific cells)
CCR7	Tm169Di	Lineage
CD25	Yb173Di	Lineage
CD56	Yb174Di	Lineage
TCRgd	Lu175Di	Lineage
CD16	Bi209Di	Lineage
CD86	In113Di	Functional (Not used in analysis)
OX40	Pr141Di	Functional (Not used in analysis)
CD28	Gd155Di	Functional (Not used in analysis)
GPR15	Gd156Di	Functional (Not used in analysis)
PD1	Er170Di	Functional (Not used in analysis)
beta 7 integrin	Yb172Di	Functional (Not used in analysis)
CLA	Yb176Di	Functional (Not used in analysis)
CD57	Y89Di	Functional (Not used in analysis)
IL-4	Nd144Di	Functional (Not used in analysis)
IL9	Sm152Di	Functional (Not used in analysis)
IFNg	Dy161Di	Functional (Not used in analysis)
IL-17	Dy164Di	Functional (Not used in analysis)
LAP	Er166Di	Functional (Not used in analysis)

IL-10	Yb171Di	Functional (Not used in analysis)
* Only Lineage markers were subjected to CyAnno for model training and cell type identification.		

References

Abdelaal,T. *et al.* (2019) Predicting Cell Populations in Single Cell Mass Cytometry Data. *Cytom. Part A*, **95**, 769–781.

Chinthrajah,R.S. *et al.* (2019) Sustained outcomes in oral immunotherapy for peanut allergy (POISED study): a large, randomised, double-blind, placebo-controlled, phase 2 study. *Lancet*, **394**, 1437–1449.

Fuss,I.J. *et al.* (2009) Isolation of whole mononuclear Cells from peripheral blood and cord blood. *Curr. Protoc. Immunol.*

Hsu,K.W. (2017) A theoretical analysis of why hybrid ensembles work. *Comput. Intell. Neurosci.*, **2017**.

Levine,J.H. *et al.* (2015) Data-Driven Phenotypic Dissection of AML Reveals Progenitor-like Cells that Correlate with Prognosis. *Cell*, **162**, 184–197.

Lin,D. *et al.* (2015) Intracellular Cytokine Staining on PBMCs Using CyTOFTM Mass Cytometry. *Bio-Protocol*, **5**.

Nassar,A. *et al.* (2015) … a Novel Technology Mass Cytometry Time-of-Flight Mass Spectrometer (Cytof2) for High-Speed Acquisition of Highly Multi-Parametric Single Cell Data: a Status …. In, *30th Congress of the ….*

Neeland,M.R. *et al.* (2020) Mass cytometry reveals cellular fingerprint associated with IgE+ peanut tolerance and allergy in early life. *Nat. Commun.*, **11**, 1–10.

Ranawana,R. and Palade,V. (2016) Multi-Classifier Systems: Review and a roadmap for developers. *Int. J. Hybrid Intell. Syst.*, **3**, 35–61.

Samusik,N. *et al.* (2016) Automated mapping of phenotype space with single-cell data. *Nat. Methods*, **13**, 493–496.

Spidlen,J. *et al.* (2012) FlowRepository: A resource of annotated flow cytometry datasets associated with peer-reviewed publications. *Cytom. Part A*, **81 A**, 727–731.

Weber,L.M. and Soneson,C. (2019) HDCytoData: Collection of high-dimensional cytometry benchmark datasets in Bioconductor object formats. *F1000Research*, **8**, 1459.

Supplementary Figures

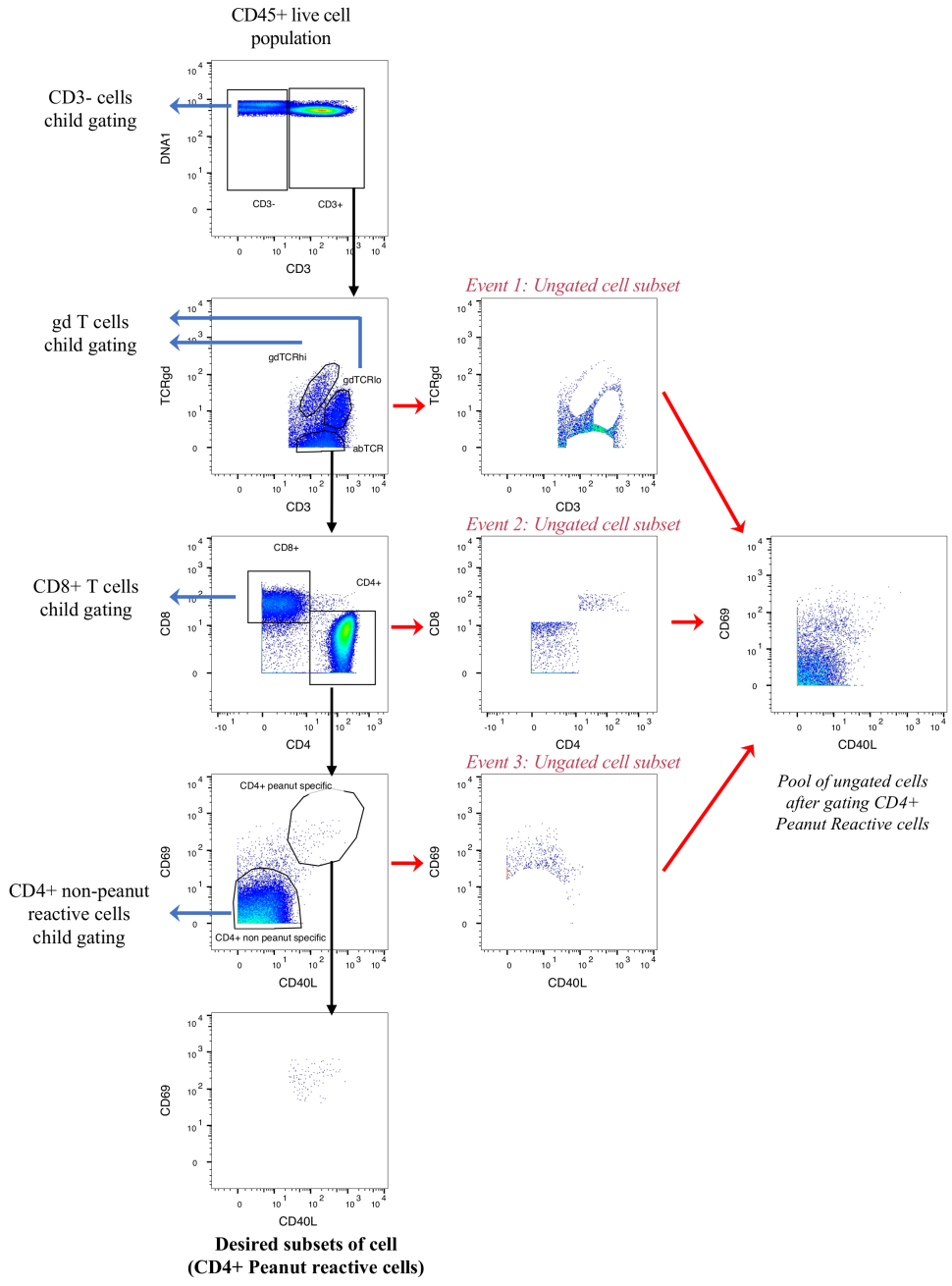

Fig. S1. Manual gating of CD4+ Peanut reactive cell type. The hierarchical schema illustrates the process involved in identification and gating of CD4+ Peanut reactive cells. Each gating event discards a population of cells as undefined, i.e. ungated. It is to be noted that each gating event can further lead to a hierarchical series of downstream child-gating for identification of other cell types (highlighted on the left). Therefore, once all the mutually exclusive cell types are identified, the cells that are ungated for all cell types together form a major population composed of heterogeneous pool of live cells, that are expected to be detected as ungated class of cells by (semi-)automated approaches.

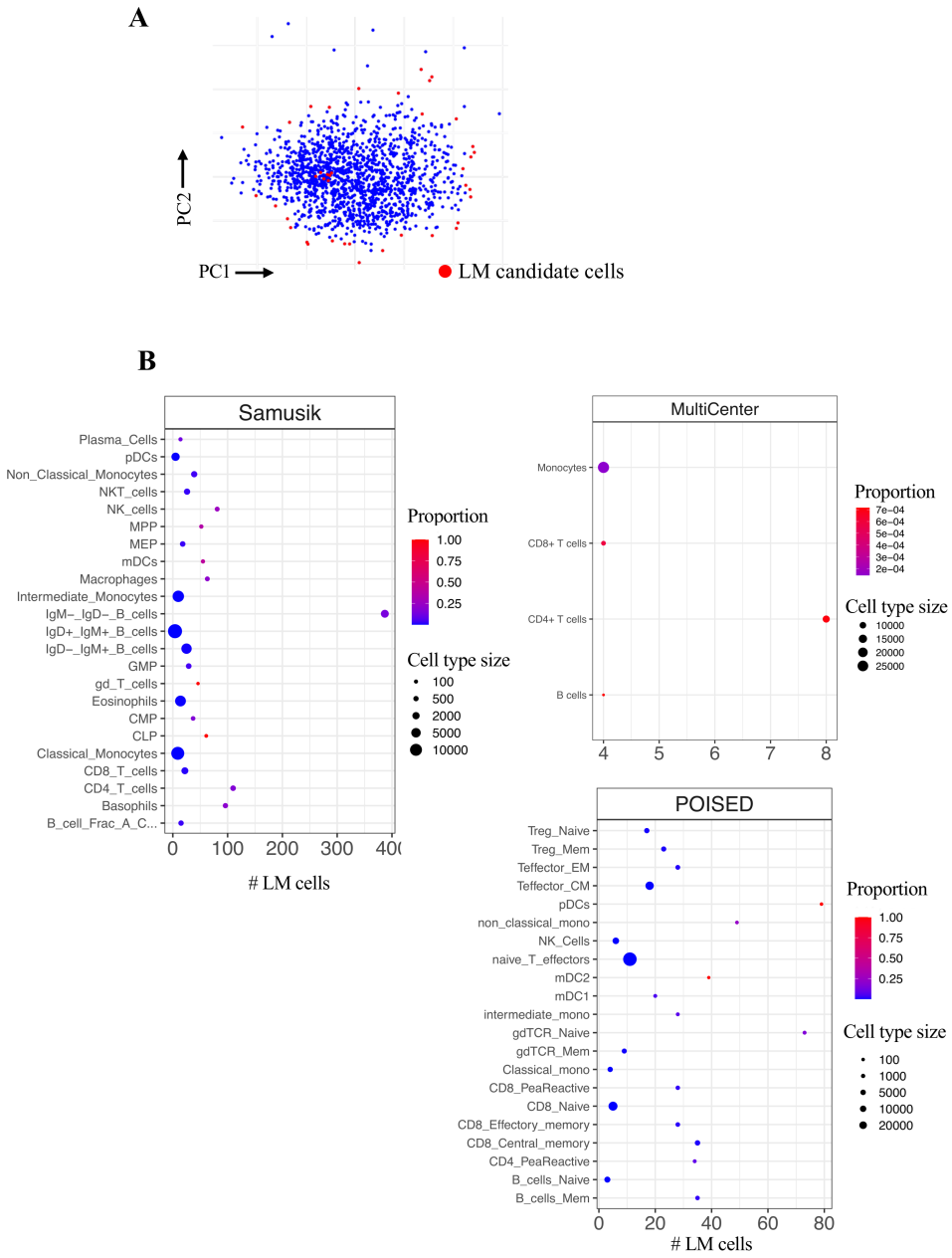

Fig. S2. LandMark (LM) cells. **A**. Schematic representation of LM cells for a cell type. Each blue dot represents a cell of a given cell type with its marker expression profile decomposed in a 2D space computed by principal component analysis (PC1 and PC2). The red dots represent LM candidates shortlisted for further screening. The LM cells on the edges represent diverse candidate cells computed via alpha shapes (see methods). **B.** The number of unique LM cells in each cell type, in each dataset used in this study. The proportion represents the number of shortlisted LM cells with respects to the total number of cells present in the training dataset for a given cell type (i.e., cell type size). Since the process of LM cell selection is stochastic in nature, the number of LM cells selected may vary during the individual runs of CyAnno for the same dataset.

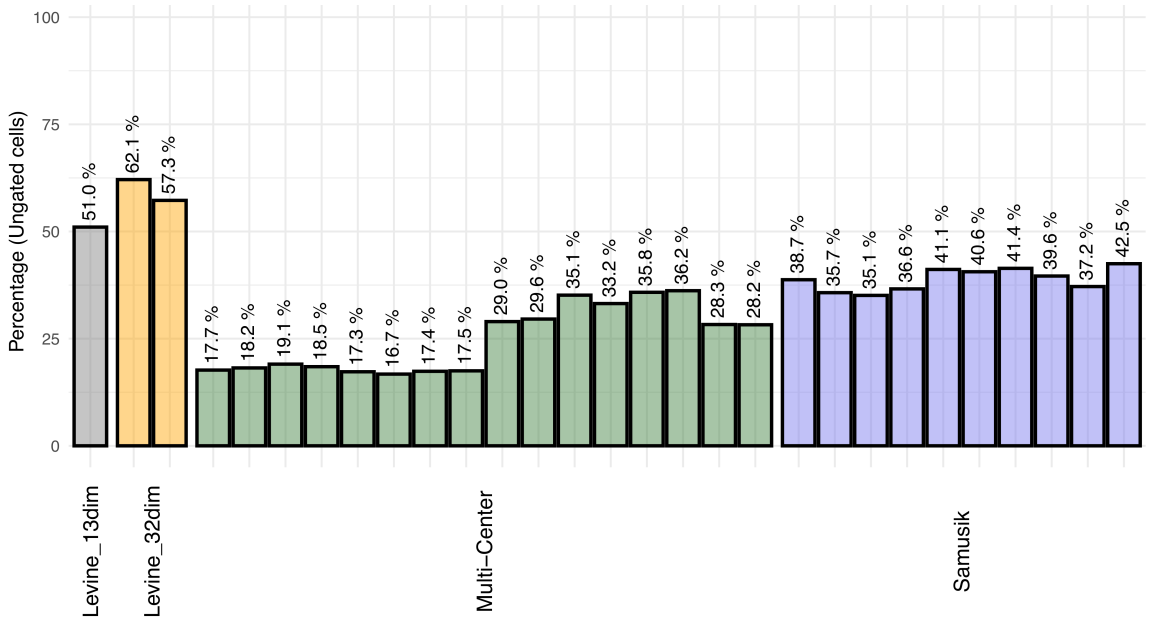

Fig. S3. Percentage of ungated cells within live cells, after manual gating, in each sample of the 4 public datasets.

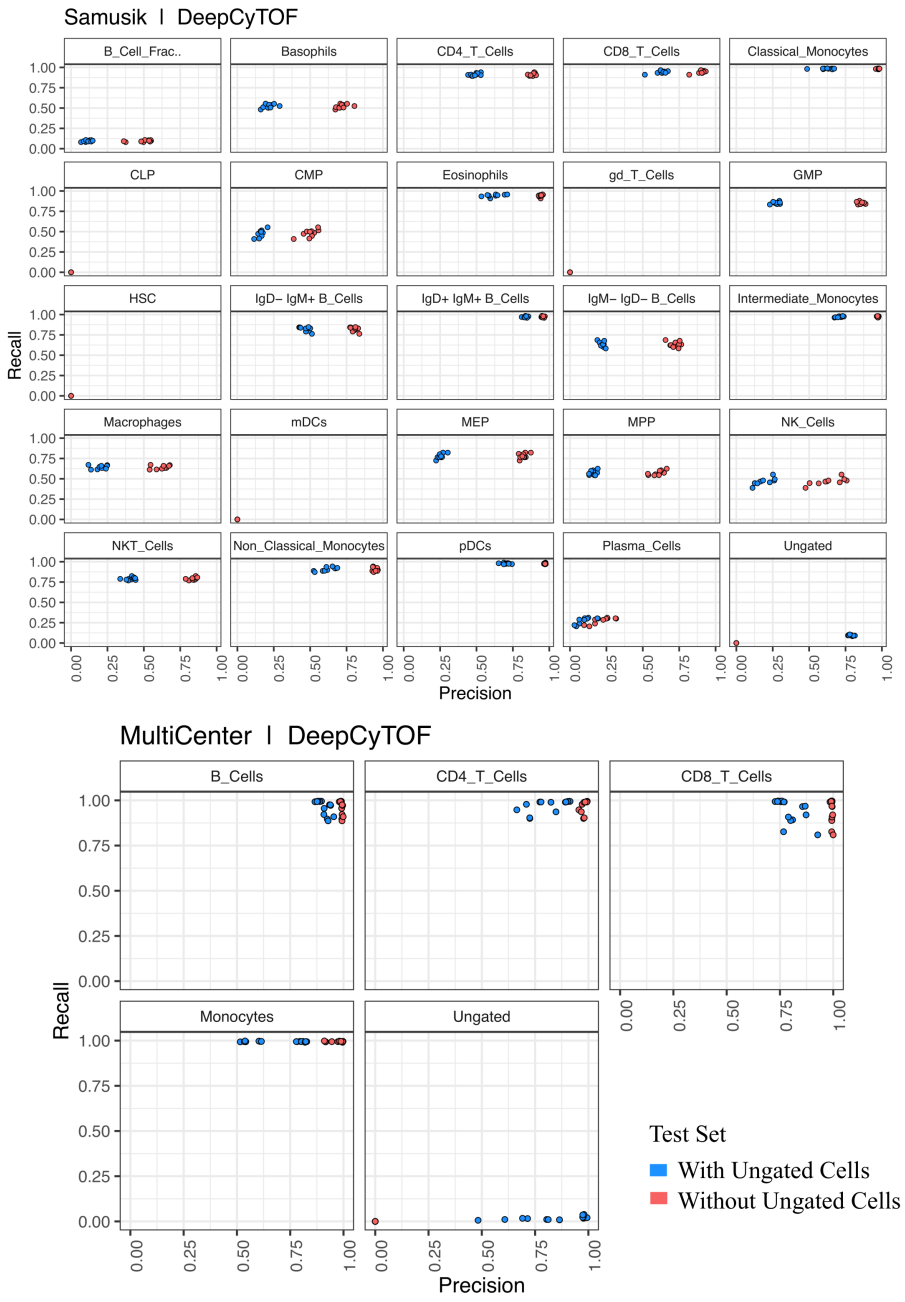

Fig. S4A. DeepCyTOF. The change in precision and recall score of prediction if ungated cells were excluded from or included in the test dataset. The precision vs recall rate shows that the models' (DeepCyTOF) ability to classify gated cells in each sample decreases when ungated cells are taken into consideration. When ungated cells were included in testing set, for most of the cell types, we observed a low precision rate, which suggests that most of these ungated cells are misclassified to one of the cell types.

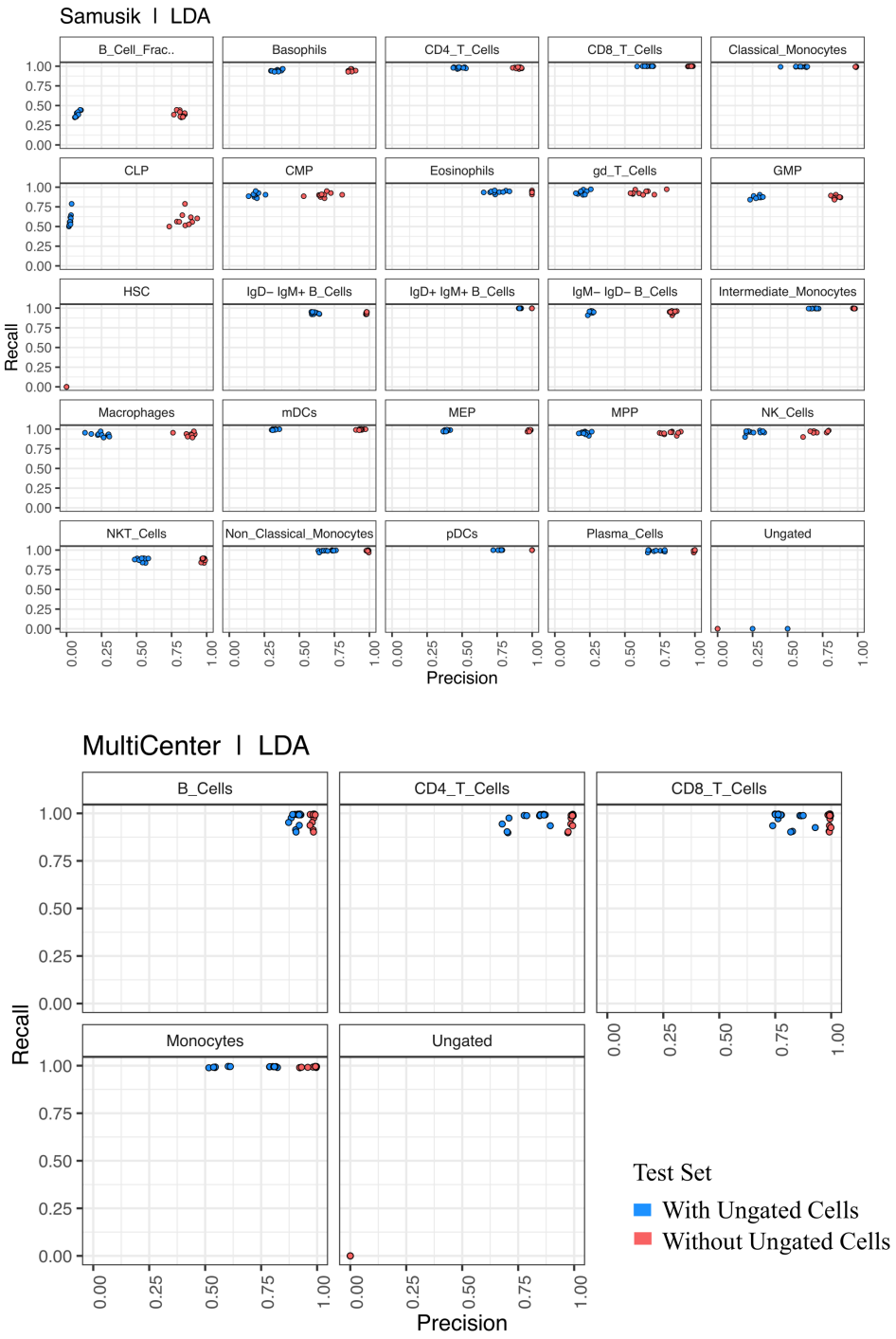

Fig. S4B. LDA. The change in precision and recall score of prediction if ungated cells were excluded from or included in the testing dataset. The precision vs recall rate shows that the models' (LDA) ability to classify gated cells in each sample decreases when ungated cells are taken into consideration. When ungated cells were included in testing set, for most of the cell types, we observed a low precision rate, which suggests that most of these ungated cells are misclassified to one of the cell types.

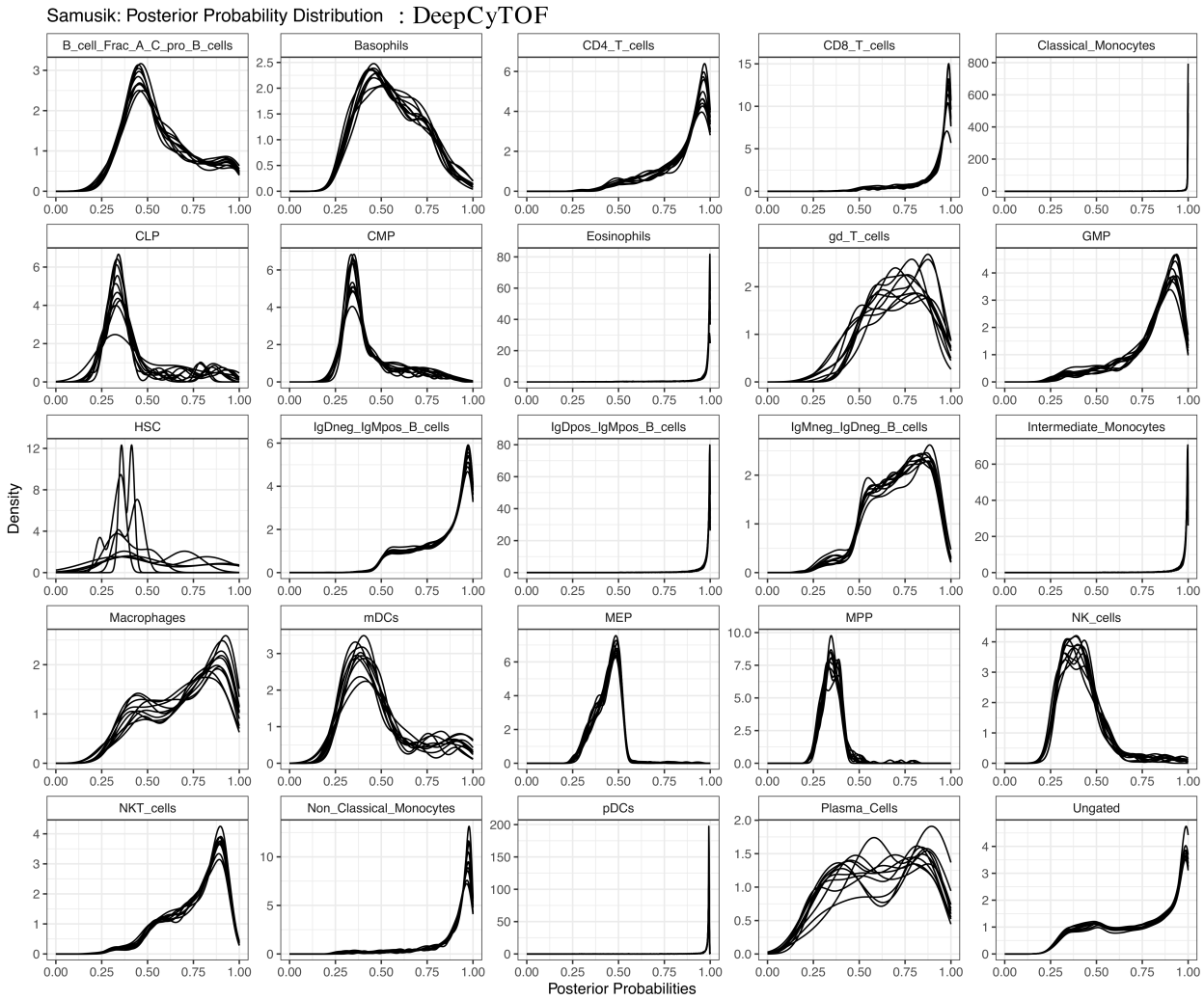

Fig. S5A. DeepCyTOF on Samusik dataset. For each cell type, the posterior probability distribution observed for cells labeled as the respective cell type after subjecting the dataset to DeepCyTOF.

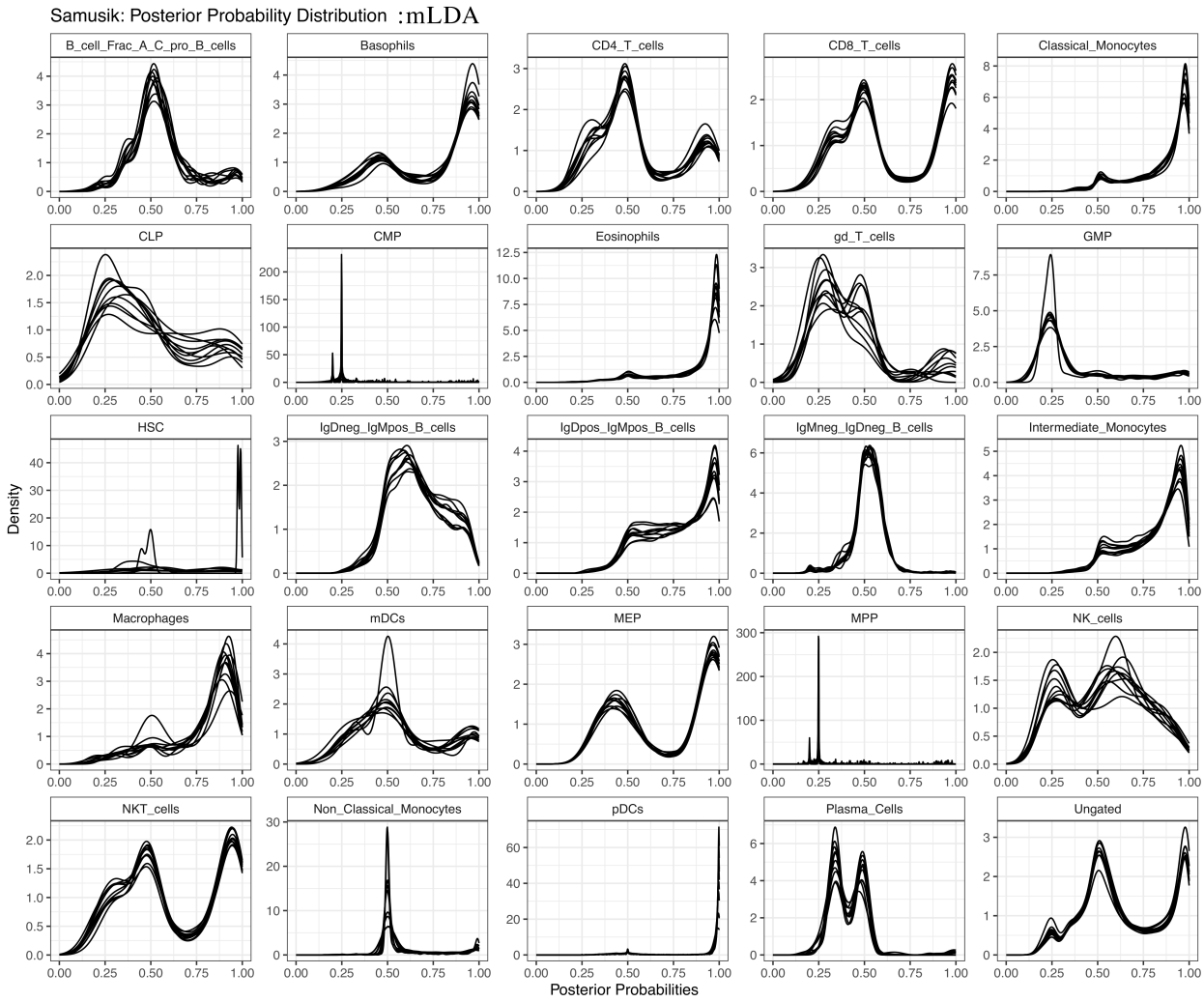

Fig. S5B. LDA on Samusik dataset. For each cell type, the posterior probability distribution observed for original cell labels after subjecting the dataset to LDA.

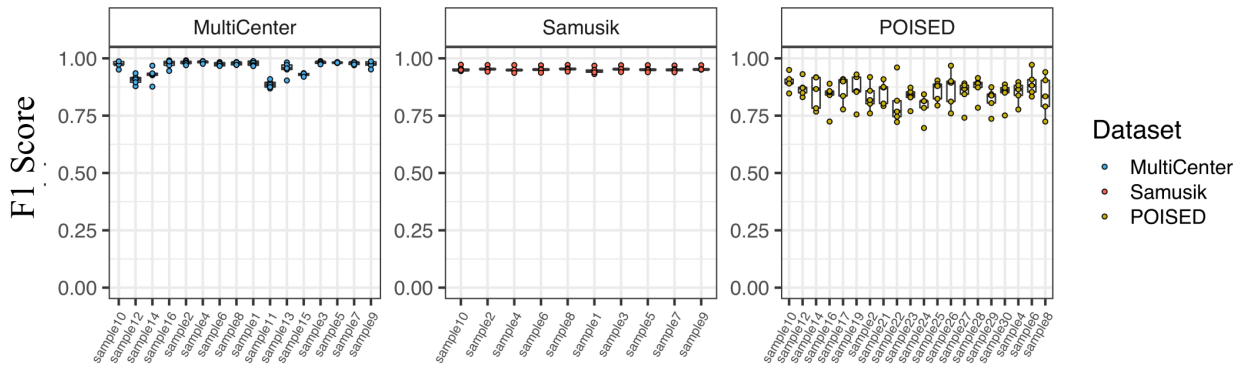

Fig. S6. F1 Score variability in each sample after 5 different runs with CyAnno. Each run involves varying training samples. For POISED dataset, training and test sets are completely independent, mixed with samples from 7 different batches and 2 different stimulations.

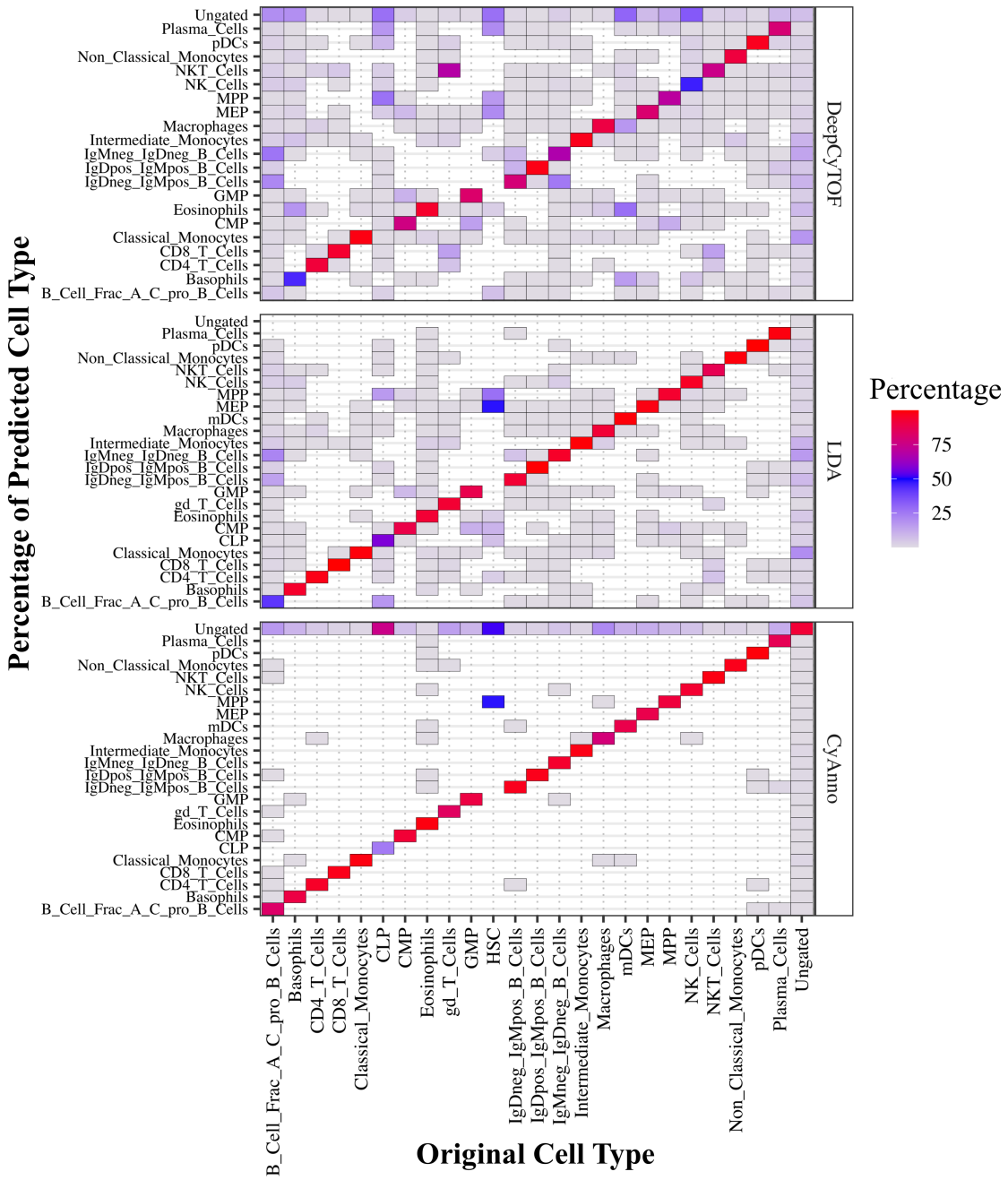

Fig. S7A. Samusik dataset. For all the cells in each of the original cell types, the composition (as percentage) of different cell type labels predicted by the models is shown. Ideally, for each cell type, an algorithm should predict low proportion of FP cells with no or minimum number of cells from other gated cell types. The different cell type labels predicted for a given original cell type suggests that most of the FP prediction in CyAnno corresponds to ungated class of cells. Here, 'Red' tiles represent high percentage of corresponding cell type labels predicted for a given original cell type, whereas blue represent low and white represent that the corresponding cell type label is not predicted for any of the cells from the original cell type.

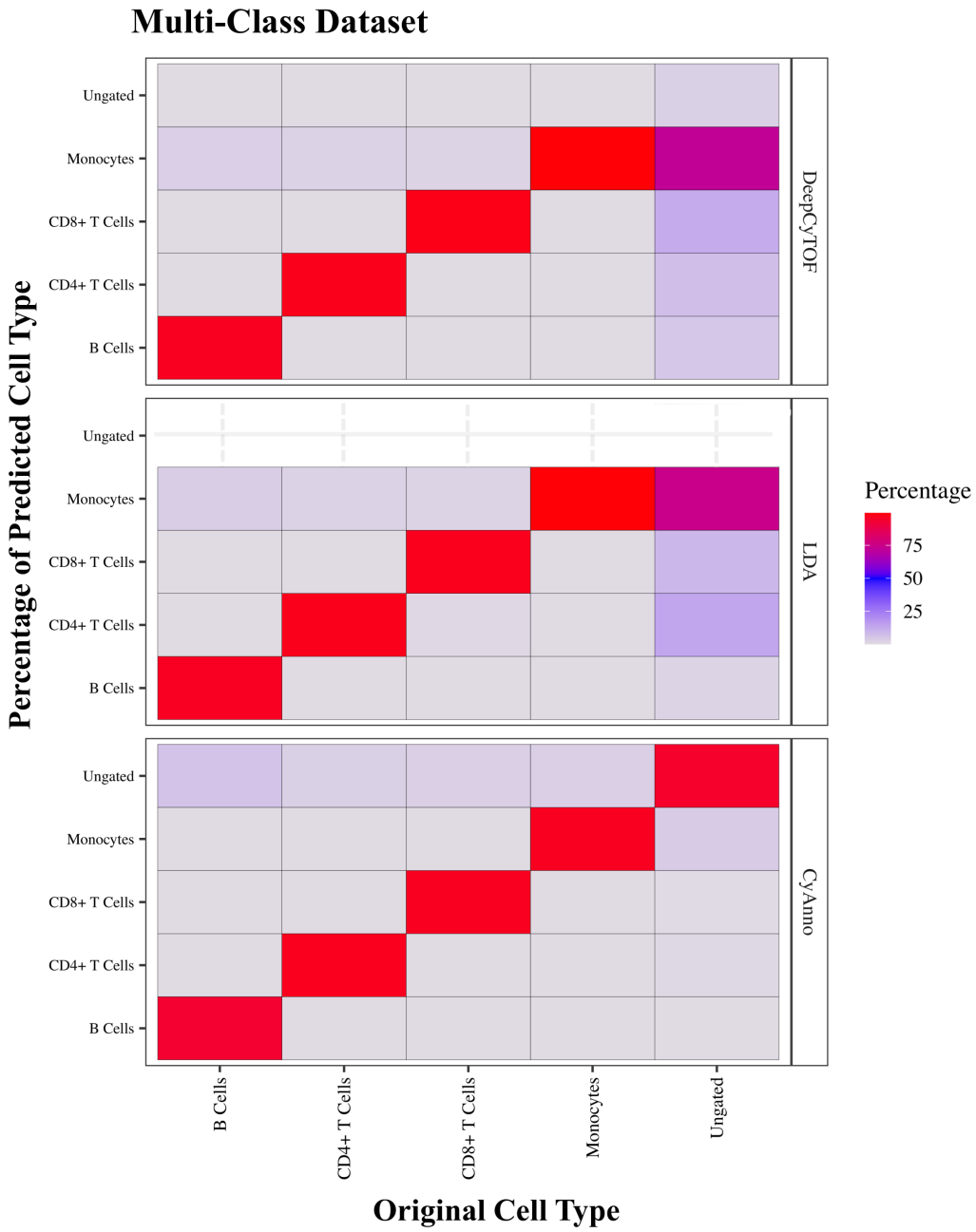

Fig. S7B. Multi-Class dataset. For all the cells in each of the original cell types, the composition (as percentage) of different cell type labels predicted by models. The composition of different cell type labels predicted for a given cell type reflects large FP prediction. Ideally, for each cell type, an algorithm should predict low proportion of FP cells with no or minimum number of cells from other gated cell types. Red boxes represent high percentage of corresponding cell type labels predicted for a given cell type, whereas blue represent low and white represent that the corresponding cell type is not predicted for any of the cells from original cell type.

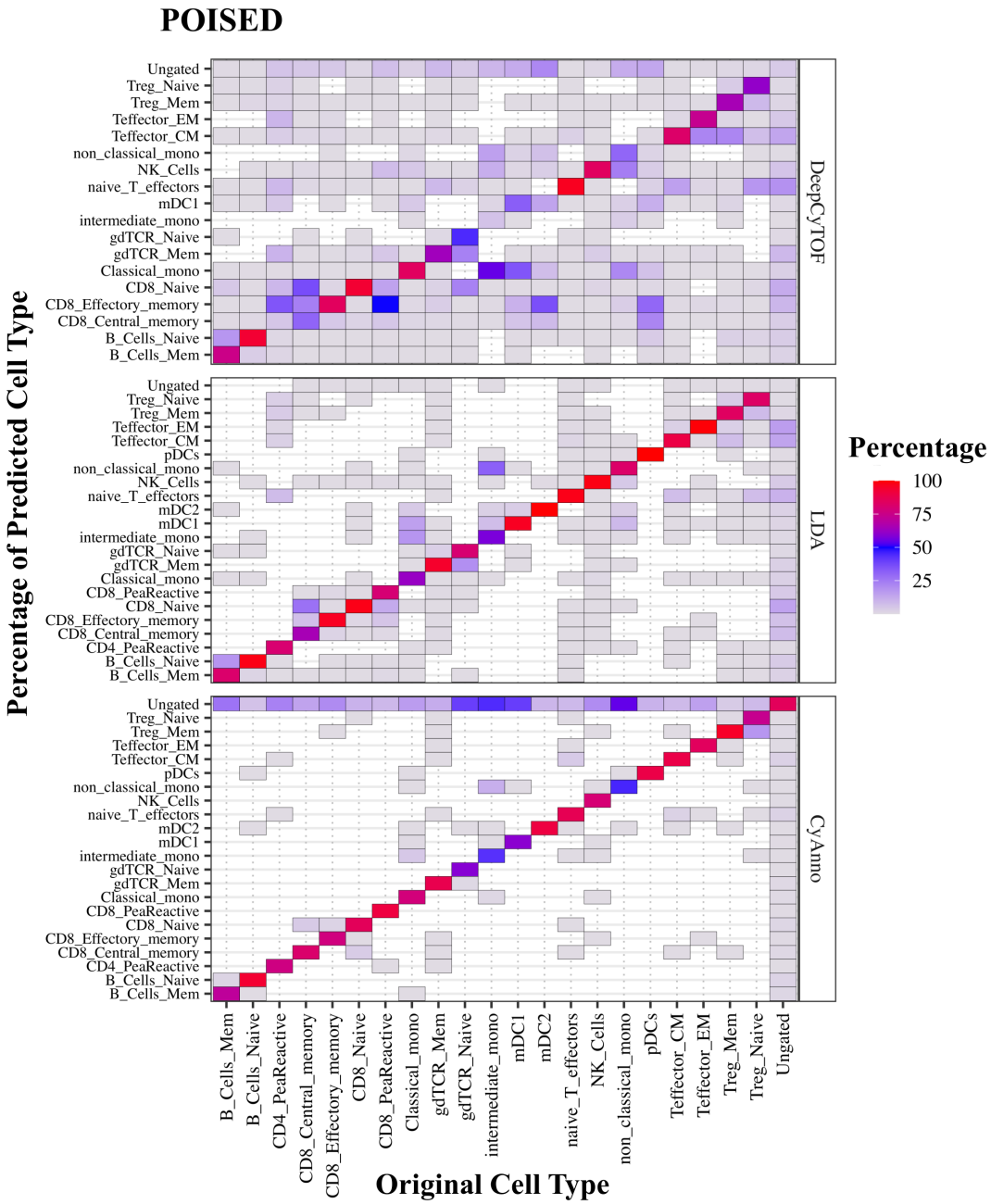

Fig. S7C. POISED dataset. For all the cells in a given cell type, the composition (as percentage) of different cell type labels predicted by models. The composition of different cell type labels predicted for a given cell type reflects large FP prediction. Ideally, for each cell type, an algorithm should predict low proportion of FP cells with no or minimum number of cells from other gated cell types. Red boxes represent high percentage of corresponding cell type labels predicted for a given cell type, whereas blue represent low and white represent that the corresponding cell type is not predicted for any of the cells from original cell type.

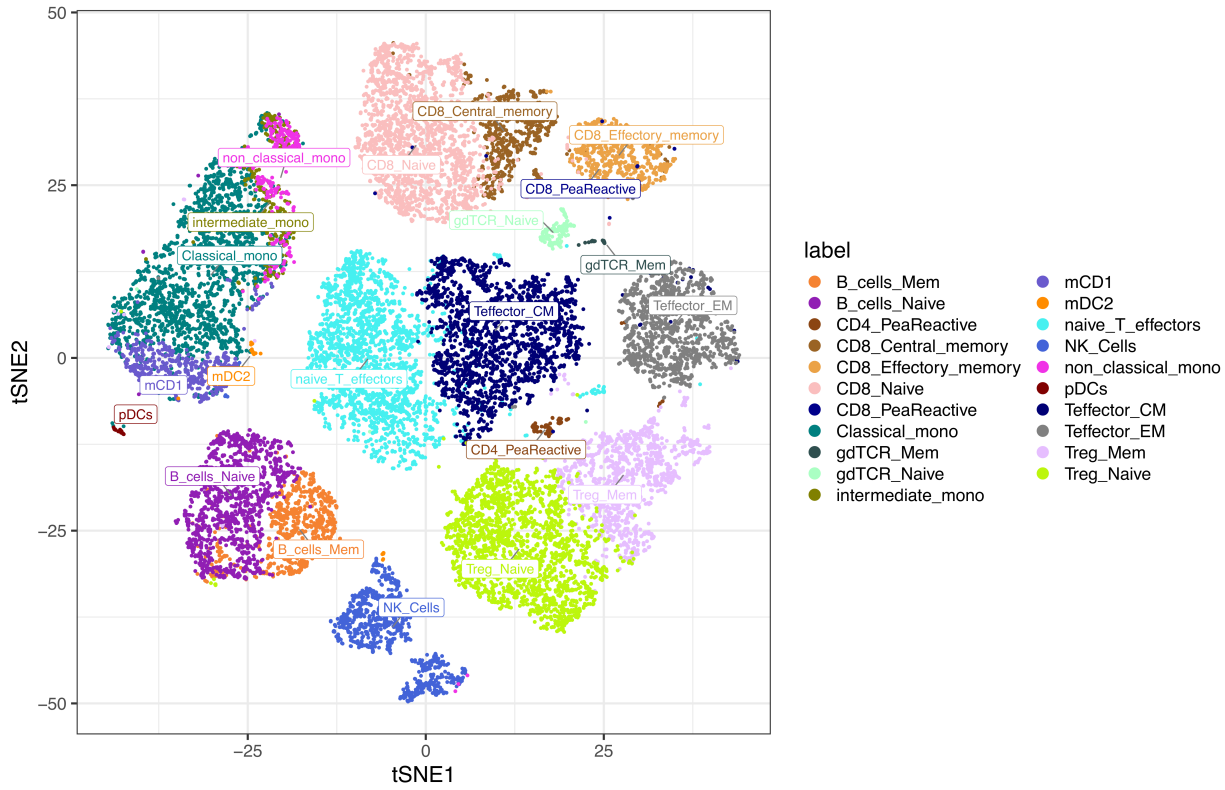

Fig. S8A. POISED dataset. tSNE visualization for POISED dataset from one sample. Maximum of 2000 cells per cell type was used for plotting the different hand-gated cell types used in this study. For clarity, ungated cells were removed from the dataset to highlight rare/small group of gated cells, e.g., peanut reactive cells (i.e., CD4+ PeaRective and CD8+ PeaReactive). It is observed that even certain gated cell types are difficult to distinguish and share non-linear classification boundaries in a 2D space, e.g., CD8+ Peanut reactive T cells or Memory B cells.

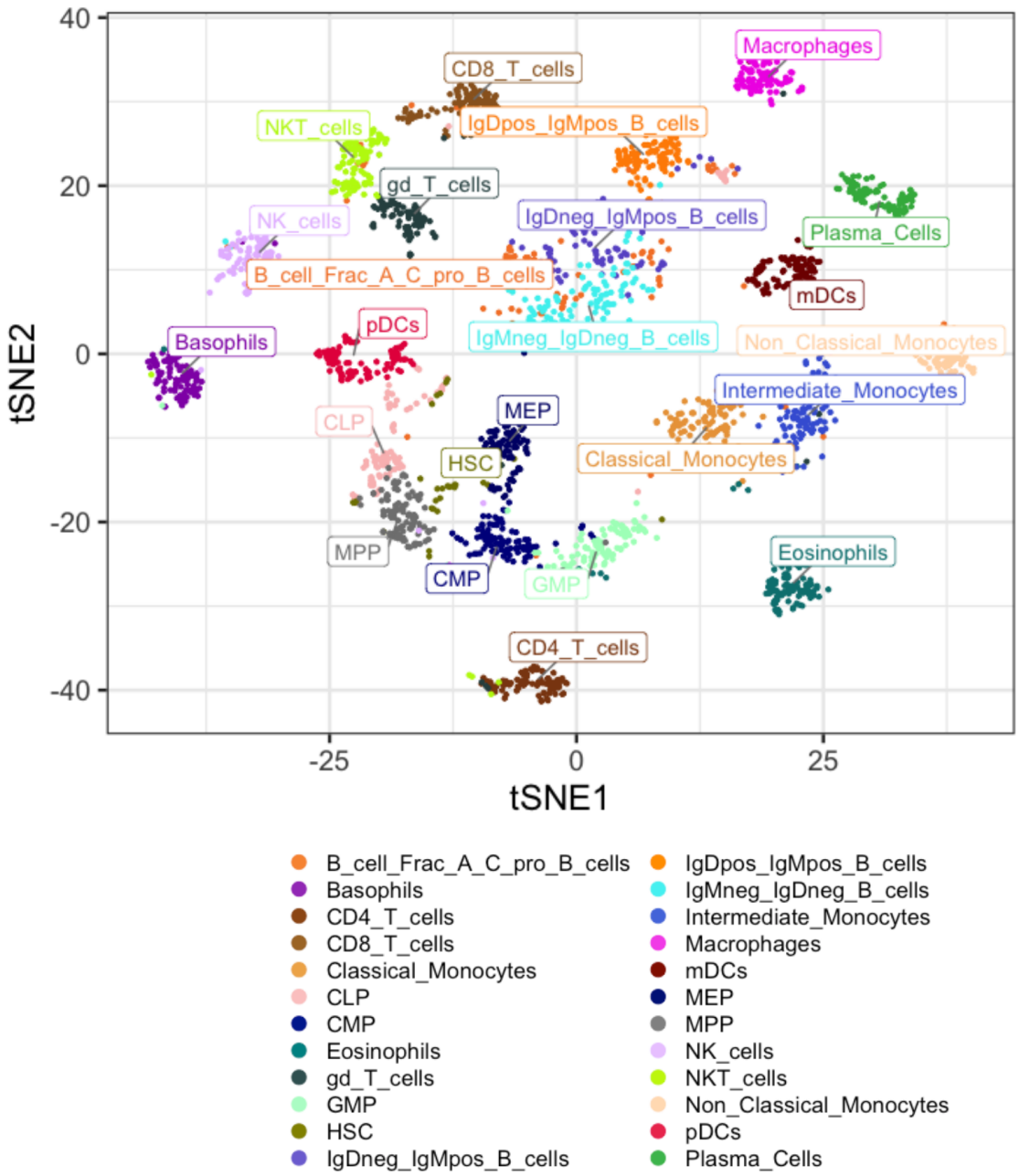

Fig. S8B. Samusik dataset. tSNE visualization for Samusik dataset. Maximum of 2000 cells per cell type, from all the samples, was used for plotting the different hand-gated cell types used in this study. For clarity, ungated cells were removed from the dataset to highlight rare/small group of gated cells, e.g., HSC.

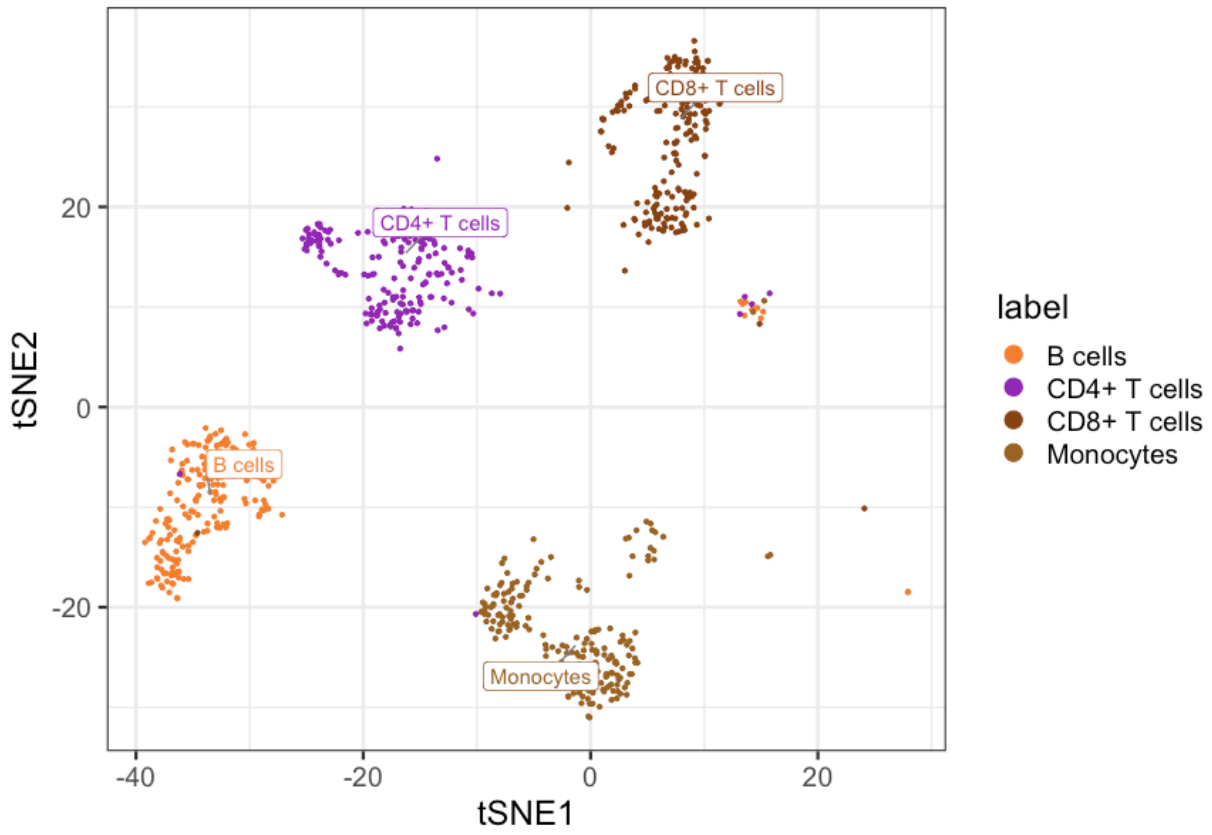

Fig. S8C. MultiCenter dataset. tSNE visualization for MultiCenter dataset. Maximum of 2000 cells per cell type, from all the samples, was used for plotting the different hand-gated cell types used in this study. For clarity, ungated cells were removed from the dataset.

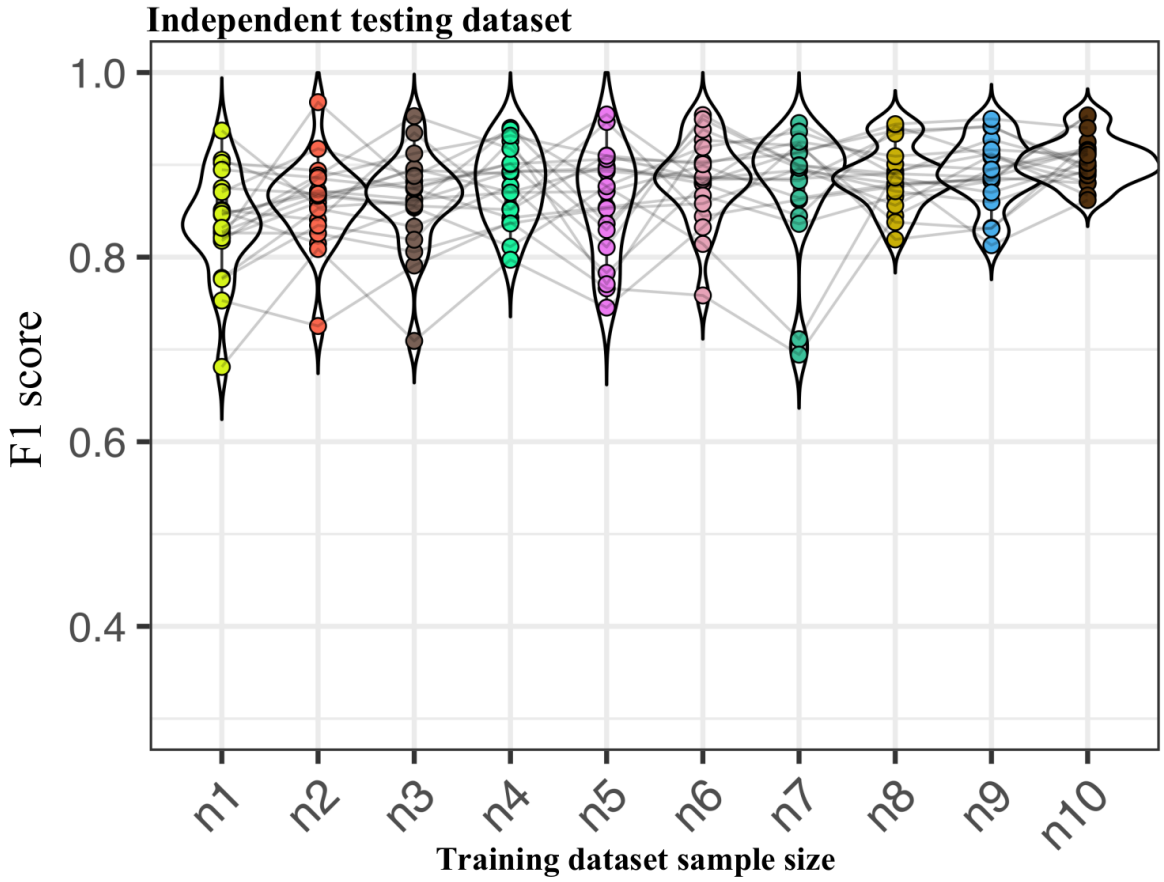

Fig. S9. CyAnno prediction with sample size variation. The F1 score per sample in independent testing set when CyAnno was executed with varying number of samples (*n*) in training set, ranging from *n*=1 to *n*=10.

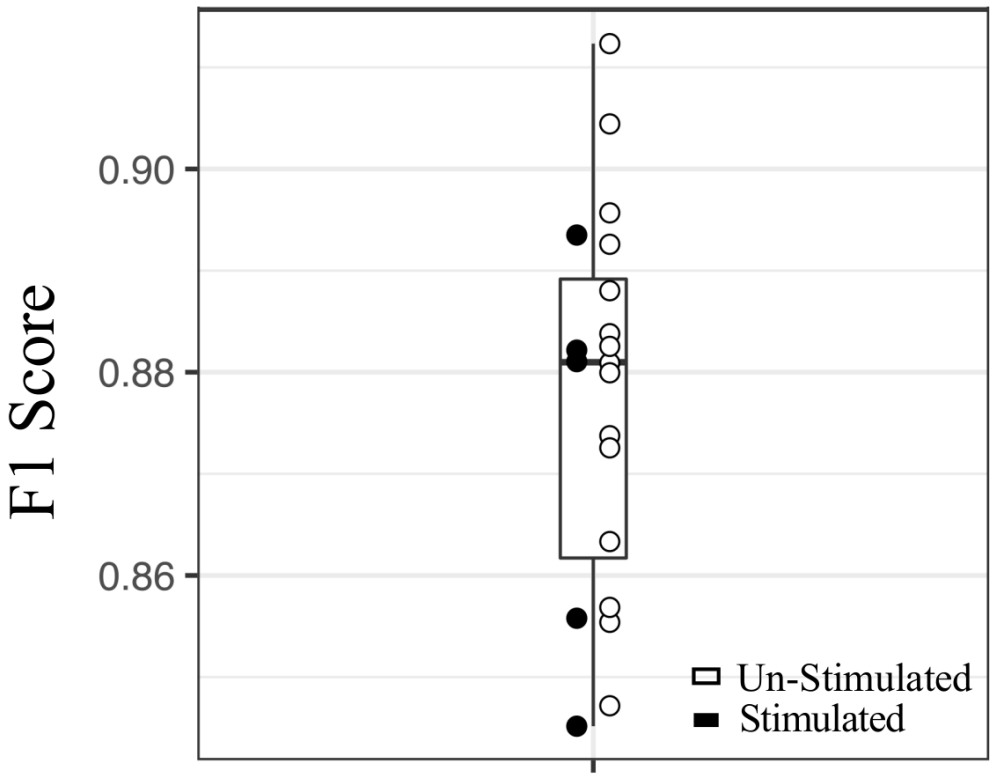

Fig. S10. The F1 score observed with CyAnno in test set samples (n=20) colored by their stimulation status with training set composed of only peanut stimulated samples.

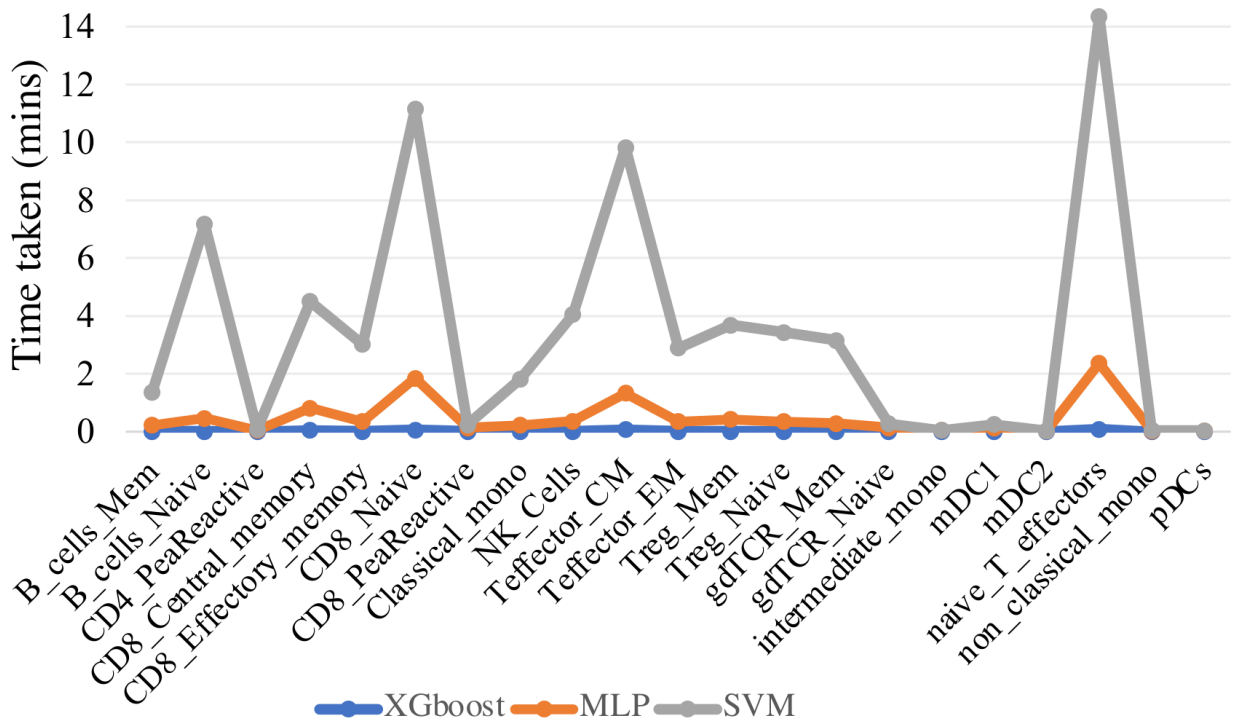

Fig. S11A. ML classifiers evaluation. Time taken by different ML classifiers used in CyAnno to train and build the hyper-parameter optimized CTSM.

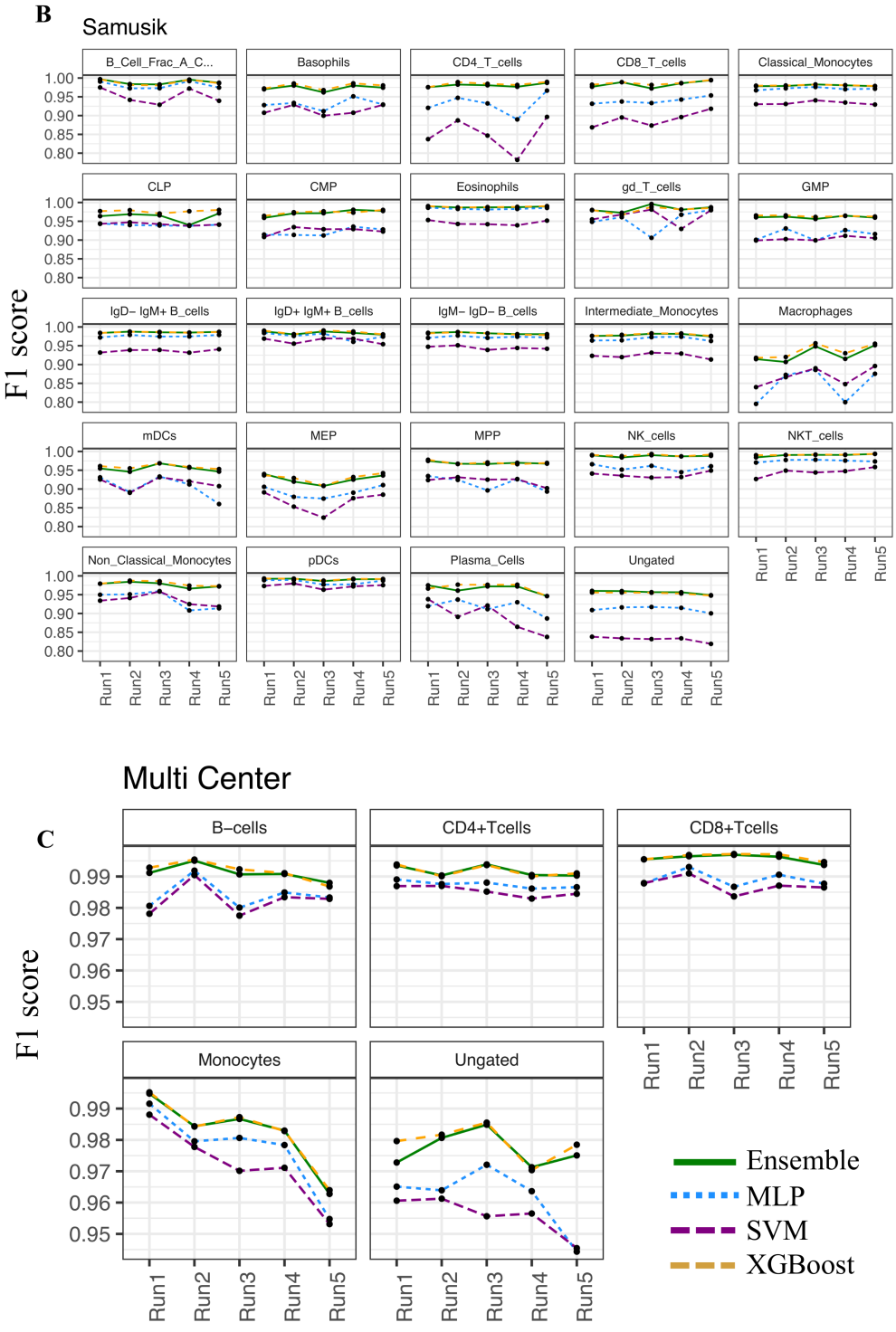

Fig. S11B-C. ML classifiers evaluation. F1 training accuracy achieved for each cell type with different classification algorithms used in CyAnno in A. Samusik dataset and B. Multi-Class dataset.

D

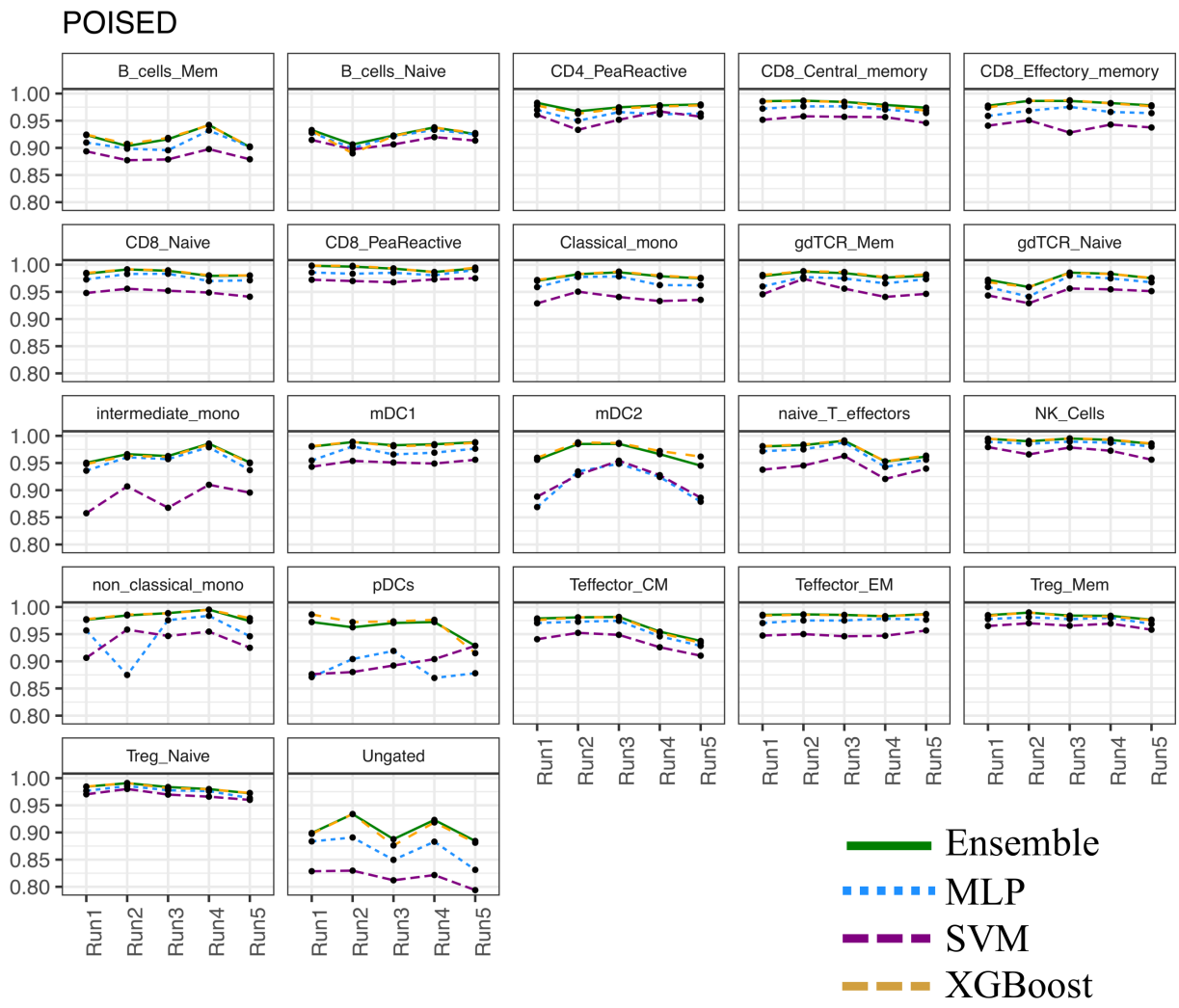

Fig. S11D. ML classifiers evaluation. F1 training accuracy achieved for each cell type with different algorithms used in CyAnno in POISED dataset.