1 Availability of evaluation data

- **Mice data set:**
  All FCS files and accompanying gates are available at `http://flowrepository.org/id/FR-FCM-ZY9G`. The files can be converted into flowLearn's data format by using the `prepare_mice.R` script provided in the GitHub repository at `https://github.com/mlux86/flowLearn`

- **FlowCAP data set:**
  We used the FlowCAP-III data available at `https://www.immunespace.org/project/HIPC/Lyoplate/begin.view`, in particular the data sets

  - `manual-gslist-bcell`
  - `manual-gslist-DC`
  - `manual-gslist-tcell`
  - `manual-gslist-treg`

  The files can be converted into flowLearn's data format by using the `prepare_flowcap.R` script provided in the GitHub repository at `https://github.com/mlux86/flowLearn`

2 Experiment Details for the Mice data set

Table 1: Experiment details for the Mice data set. The bone marrow analysis was performed on a four laser BD Fortessa X-20. For all markers, clones fluorochromes, and lasers are listed.

Marker	Clone	Fluorochrome	Laser	LP filter	BP filter
Live/Dead	na	NIR	640	750	780/60
CD45	30-F11	Qdot605	405	595	610/20
B220	RA3-6B2	PE-Cy7	561	750	780/60
CD43	1B11	PerCP-Cy5.5	488	685	710/50
CD24	M1/69	APC	640	-	670/14
BP-1/Ly51	6C3	PE	561	570	585/15
IgM	RMM-1	BV421	405	-	450/50
IgD	11-26c.2a	FITC	488	505	530/30
Gr1	RB6-8C5	AF700	640	690	730/45
CD11b	M1/70	BV510	405	505	525/50
CD3	17A2	BV786	405	750	780/60
CD138	281-2	BV650	405	630	670/30

3 Sample densities colored by center

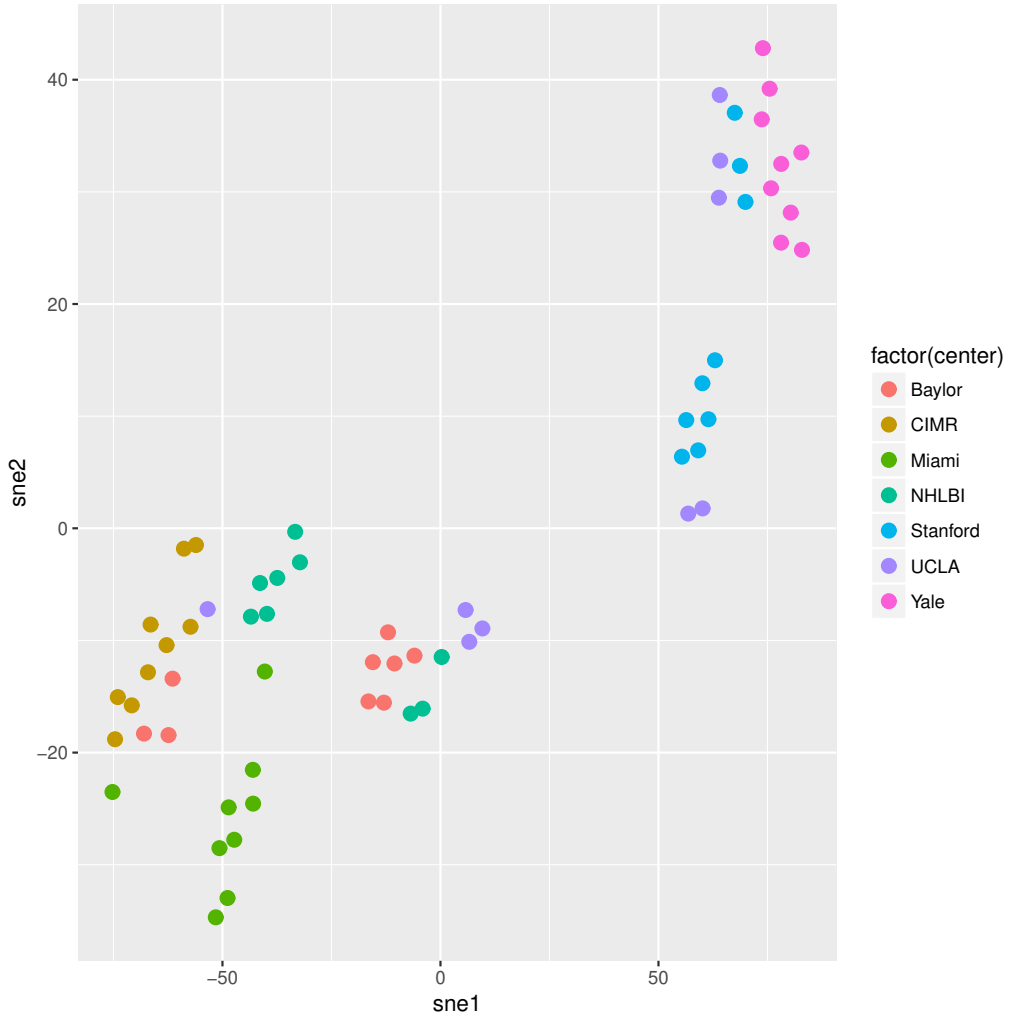


Figure 1: Exemplary illustration of the diversity in the FlowCAP data set: Densities from the T-cell CD8-activated population estimated by flowLearn were projected to two dimensions using t-SNE and colored by center. Each data point represents one density of that population. It is visible that within-center densities are more similar to each other than to other centers, indicating between-center variability, i.e. an analysis bias per center. This result is in concordance with results shown in Finak et al., 2016.

4 t-SNE analysis of the Mice data

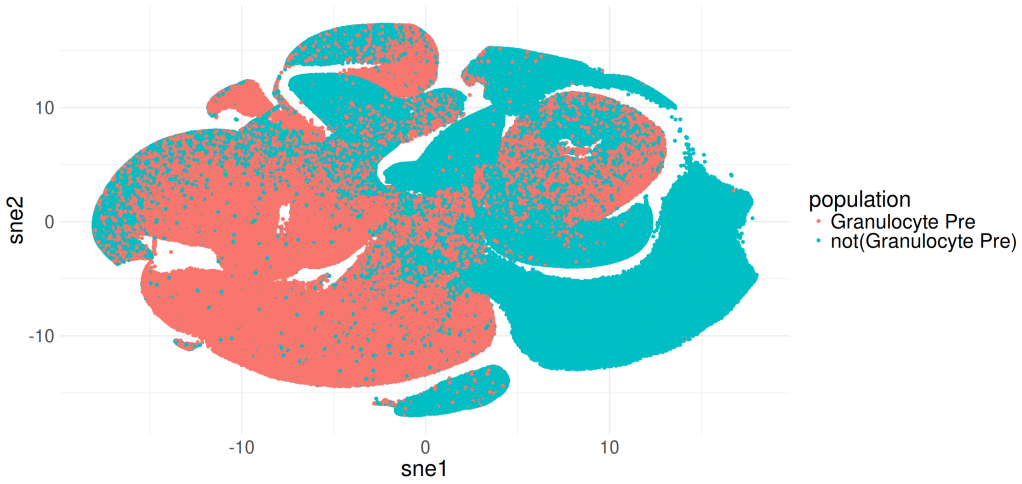


Figure 2: T-sne plot of a population of 345637 CD45 cells (dots) from one sample of the bone marrow Mice data, colored by the child populations "Granulocyte Pre" and "not(Granulocyte Pre)" cells. It is clearly visible that both populations do not form distinct clusters in this representation. Consequently, it is highly difficult to correctly separate the populations using this approach.

5 Explanation of low threshold variability in CD43+/CD43-

In the CD43+ and CD43- populations, there is a larger dispersion between true and predicted cell proportions when compared to other populations. More specifically, the true cell proportions have very low variance, when compared to other population. We got in touch with those who created the gate and it became clear that the true thresholds were strongly post-processed to meet requirements of independent analysts who verified the gate. In particular, due to overlapping clusters on the CD43 channel, it was difficult for them to find good thresholds for all samples. Therefore, many thresholds were individually and carefully modified. This resulted in an overall lower variance of the true CD43+ and CD43- cell proportions when compared to other populations. On this population, our method did not fully match all predicted cell proportions, however it is noted that performance on this population in terms of the F1 score is still excellent (median 95%), by using one prototype only, and also that the performance increases when choosing more prototypes.

6 HFC population analysis

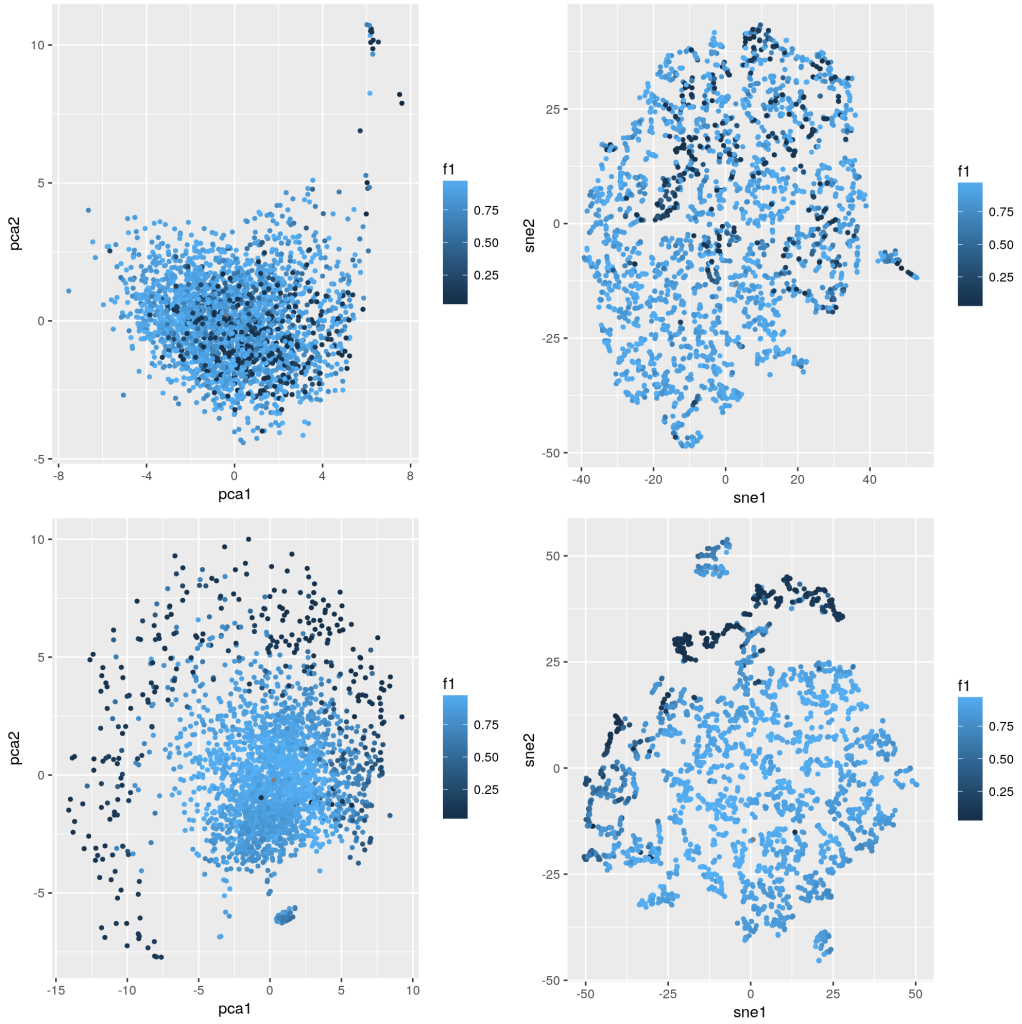


Figure 3: PCA (left) and t-SNE (right) projections of the Mice data / HFC population densities from both channels (top, bottom), colored by F1-score according to predictions using $n_p = 1$. It is visible that the second channel can be identified as the source of low performance: in both PCA and t-SNE, samples with low performance cluster together, indicating some anomalies in that channel.

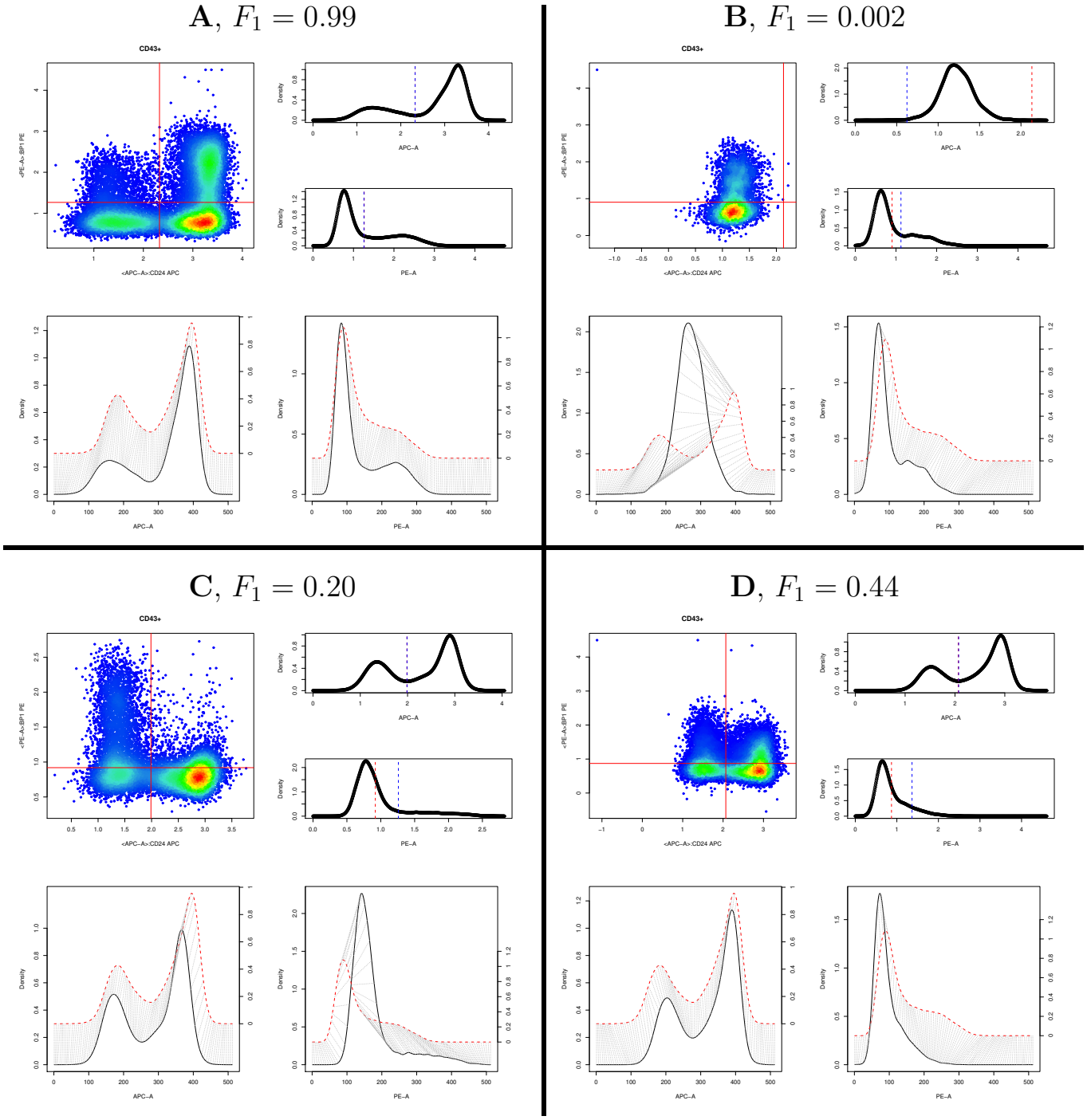


Figure 4: Analysis of four samples (A,B,C,D) of the HFC population with different prediction performances. For each sample, on the top-left a bivariate density plot is shown. The top-right shows densities for each individual channel, true (red) and predicted (blue) thresholds. Bottom-left and bottom-right plots show alignments with the prototype density for both channels, respectively. A: Normal sample with well pronounced HFC population (top right in the bivariate density). B: Missing HFC population resulting in a failed alignment of the APC-A channel. C: Missing HFC population resulting in an inaccurately set training threshold for the PE-A channel. D: Weakly pronounced HFC population and resulting wrong alignment of the PE-A channel.

7 Gate differences in the FlowCAP data set

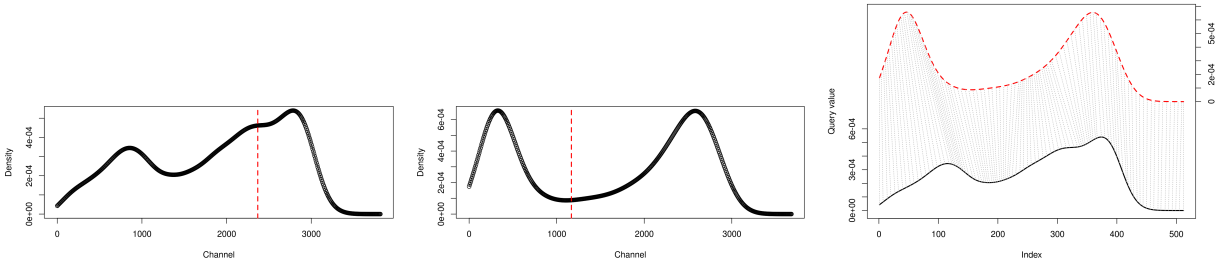


Figure 5: Differences in samples and gates in the FlowCAP data set, exemplary for one channel of the CD4 Effector population: Density plots show the population density from two samples (left, center). It is visible that the true thresholds (red) are much different from each other. The alignment between the two top densities is shown on the right. Even though the alignment looks correct, because of the difference in true thresholds, they are mispredicted by flowLearn.

8 Example of gating a rare population

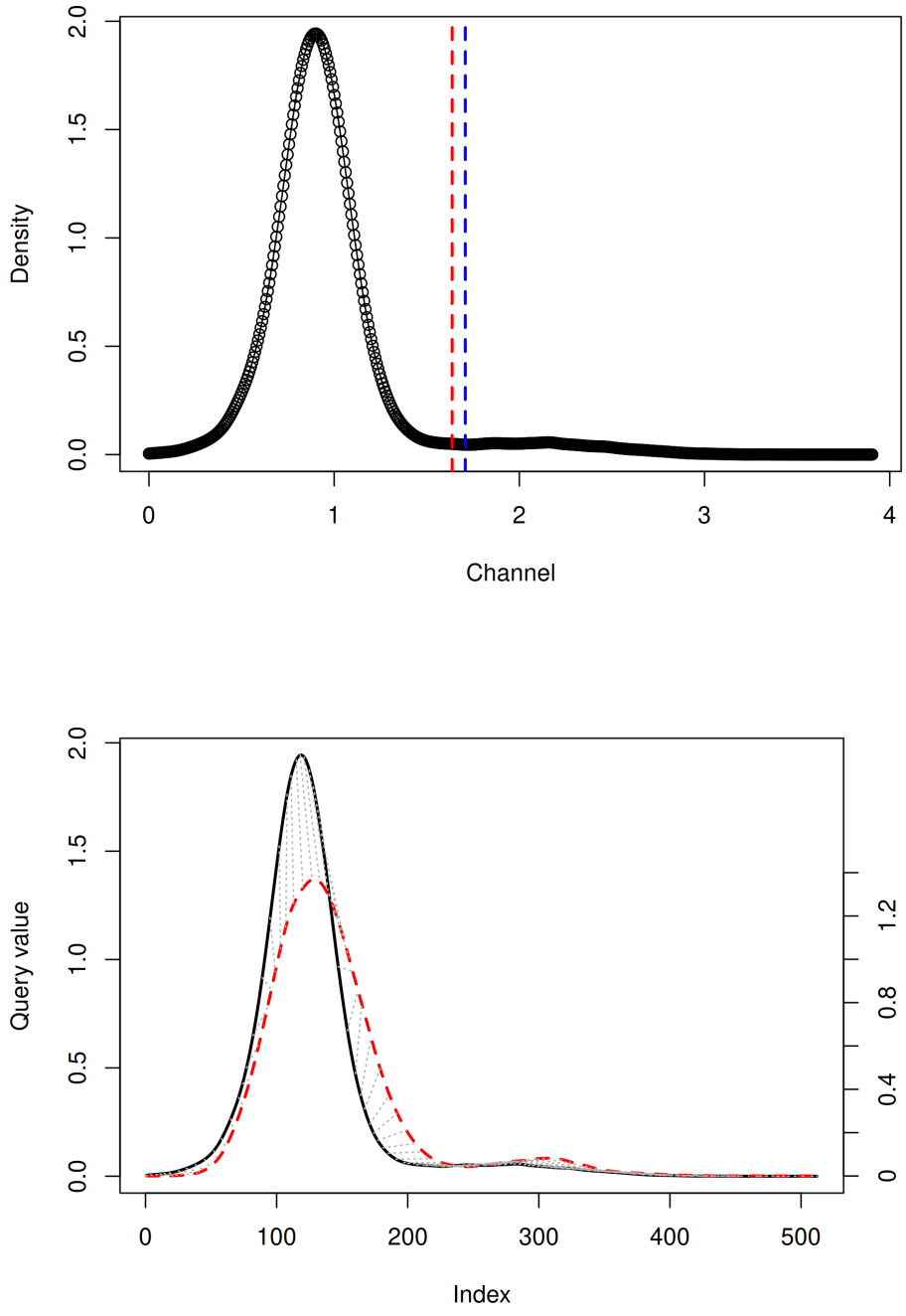


Figure 6: Example of gating a rare population (T cells, 5% of parent). Top: true (red) and predicted (blue) threshold on the parent density. The small bump on the right side of the density corresponds to the T cell population. Rare populations result in flat parts of the density, although peaks do not need to be very pronounced in order to be aligned, when there is another, larger density peak that can be aligned properly. Only a few cells of another rare population are needed in order for the density to extend into a tail, containing the rare population. That the tail is aligned correctly follows directly from the correct alignment of the adjacent, larger peak. Bottom: Alignment with of the top density (black) with a prototype (red dashed).

9 Comparison of flowLearn, FlowSOM and DeepCyTOF

FlowSOM: We ran the `FlowSOM()` function using their package `1.8.1`. It required two mandatory parameters: the set of considered channels is chosen as the set of channels with protein markers (flowLearn uses the same set for gating), and the number of clusters was chosen as the number of target populations, plus one "ungated" population, in total 11. Due to the missing population labels for each cluster, we matched populations using a posterior labeling with the Hungarian assignment algorithm. For this part, we used the same code as Weber et al. in their comparison paper. The choice of parameters, as well as the matching of populations is in concordance with the evaluation procedure used in the comparison paper by Weber et al. We calculated F_1-scores in the same way as for flowLearn (per population).
DeepCyTOF: We obtained the DeepCyTOF source code from their GitHub repository, using the commit ID `ac924b10a77cf2d5c7ea43ea39a172898d71abb1`. From the Mice data set, we generated CSV files with the same format as used by the authors. Furthermore, we modified the main script `DeepCyTOF.py` to include our data set. Out of time reasons, we used every 50th out of all 2665 samples as a potential training sample for DeepCyTOF, from which it chose one. The channels are the same that are used for FlowSOM and flowLearn (protein marker channels). The number of classes is fixed to 11. Following the procedure in the author's paper, we skipped training the auto-encoder network, because there are no missing values in our data set. Also, we did not use model calibration for each new sample. When we included it for a set of samples, it did not improve results. Since all samples were from the same flow cytometry machine, this is in concordance with results in their original paper. Furthermore, it is worth to note that F_1-scores reported by DeepCyTOF cannot directly be compared to F_1-scores from flowLearn. This is because DeepCyTOF uses a "micro"-F_1, i.e. globally counts true/false positives/negatives, whereas flowLearn calculates "macro"-F_1 values for each population, independent of its size. Hence, we exported predicted populations per cell and evaluated the F_1-score in the same manner as for flowLearn.

Table 2: Description of the evaluation procedures for both DeepCyTOF and FlowSOM.

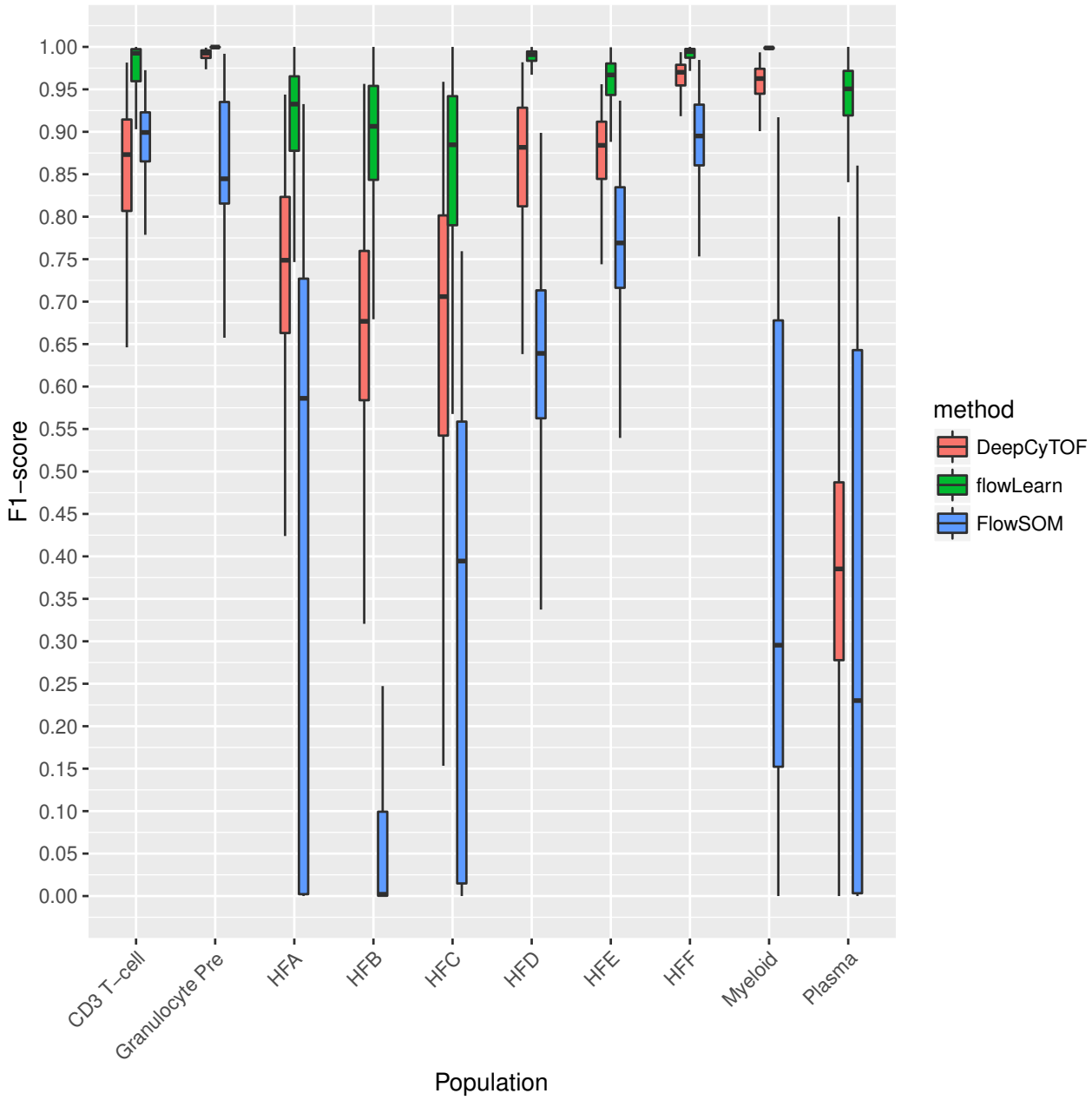


Figure 7: F_1-scores obtained by running flowLearn, DeepCyTOF, and FlowSOM on the leaf populations of the Mice data set.

10 Full results on the Mice data set

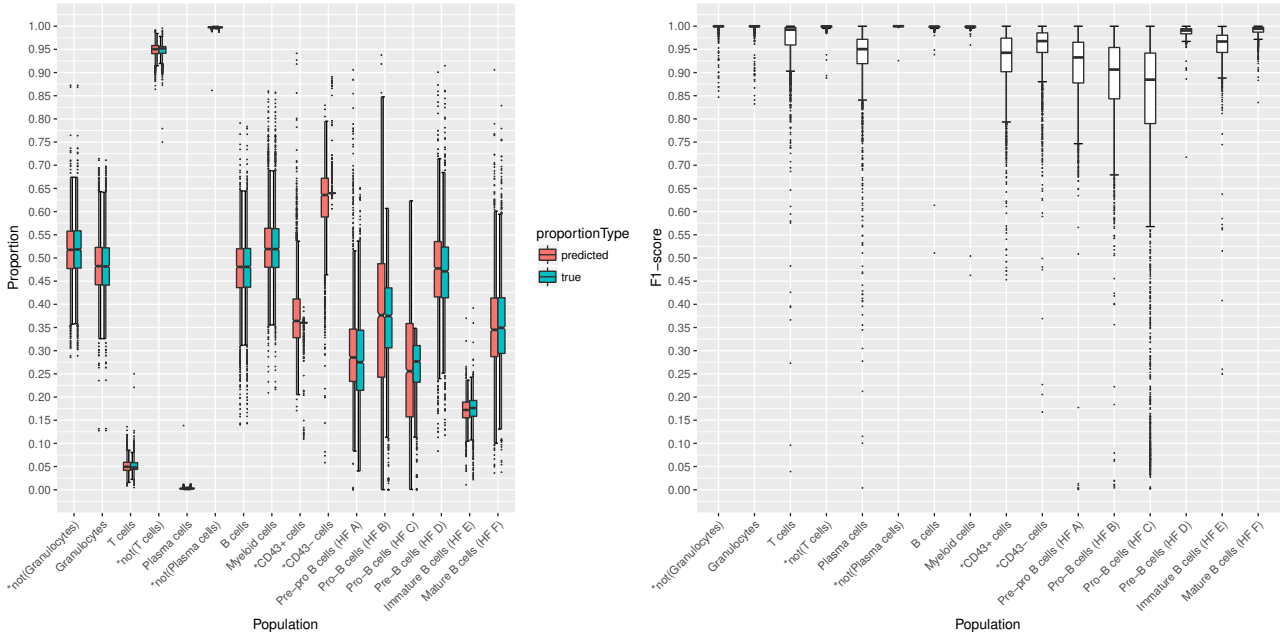


Figure 8: Results on the Mice data set for $n_p = 1$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (2665 samples). Outliers are shown as single dots. Populations HFA to HFF stand for Hardy-Fraction A to F (explained in Supplementary Table 2). Populations denoted with an asterisk are non-biological (technical) populations.

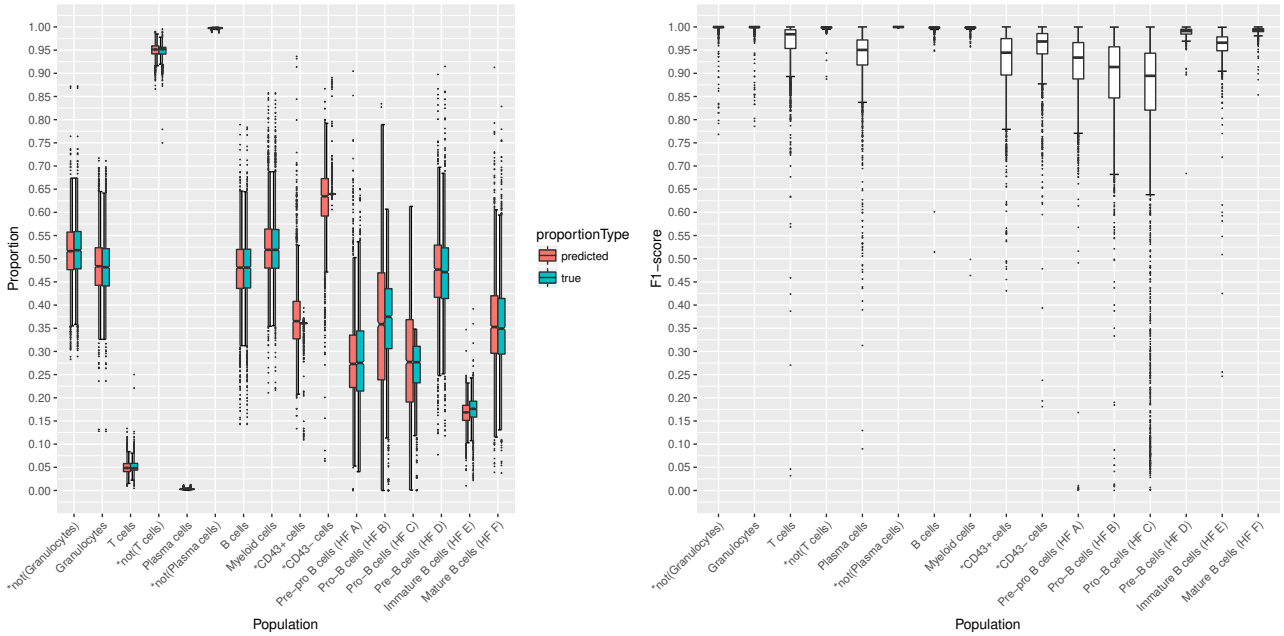


Figure 9: Results on the Mice data set for $n_p = 2$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (2665 samples). Outliers are shown as single dots. Populations HFA to HFF stand for Hardy-Fraction A to F (explained in Supplementary Table 2). Populations denoted with an asterisk are non-biological (technical) populations.

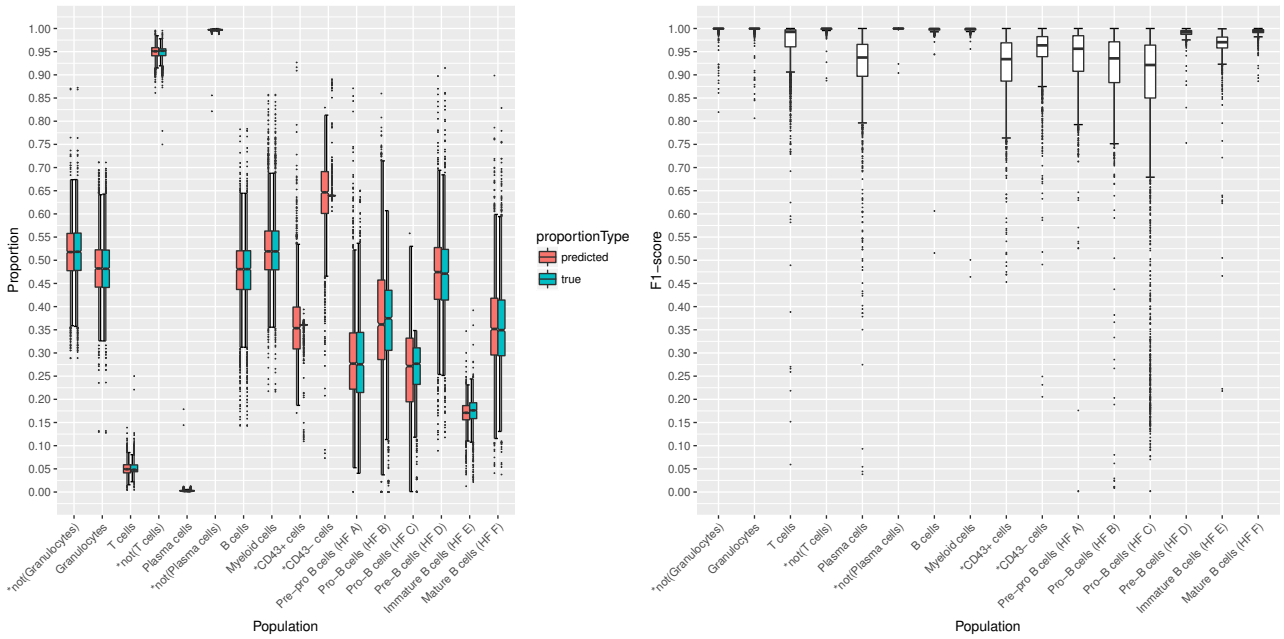


Figure 10: Results on the Mice data set for $n_p = 5$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (2665 samples). Outliers are shown as single dots. Populations HFA to HFF stand for Hardy-Fraction A to F (explained in Supplementary Table 2). Populations denoted with an asterisk are non-biological (technical) populations.

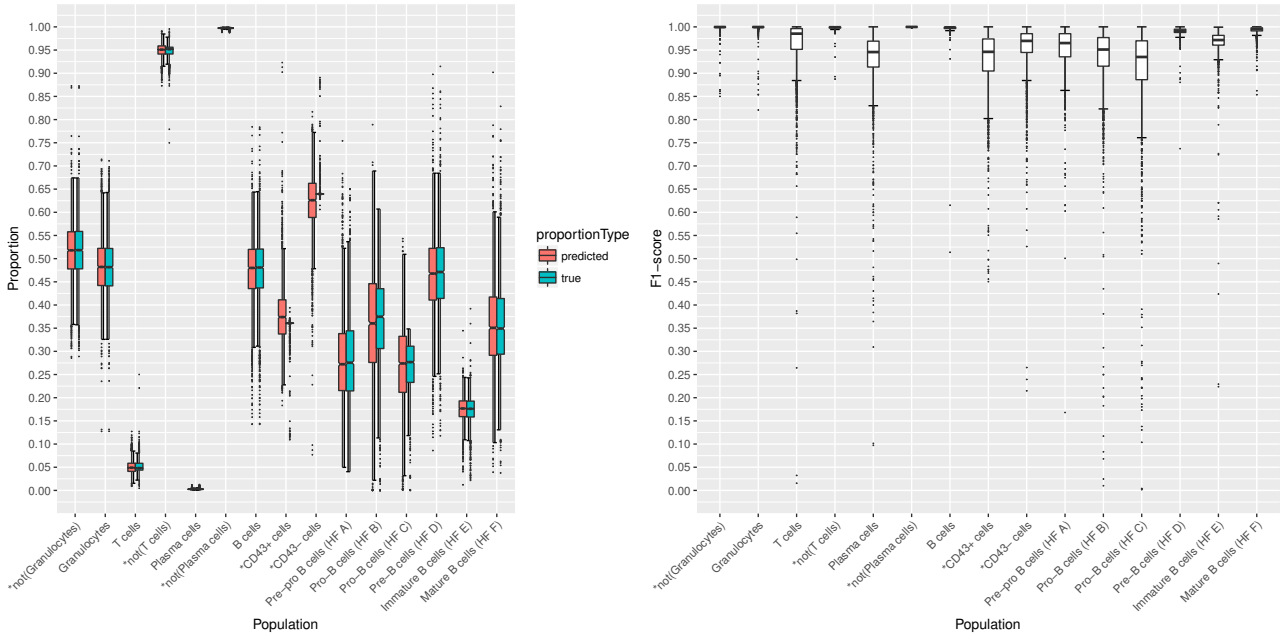


Figure 11: Results on the Mice data set for $n_p = 10$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (2665 samples). Outliers are shown as single dots. Populations HFA to HFF stand for Hardy-Fraction A to F (explained in Supplementary Table 2). Populations denoted with an asterisk are non-biological (technical) populations.

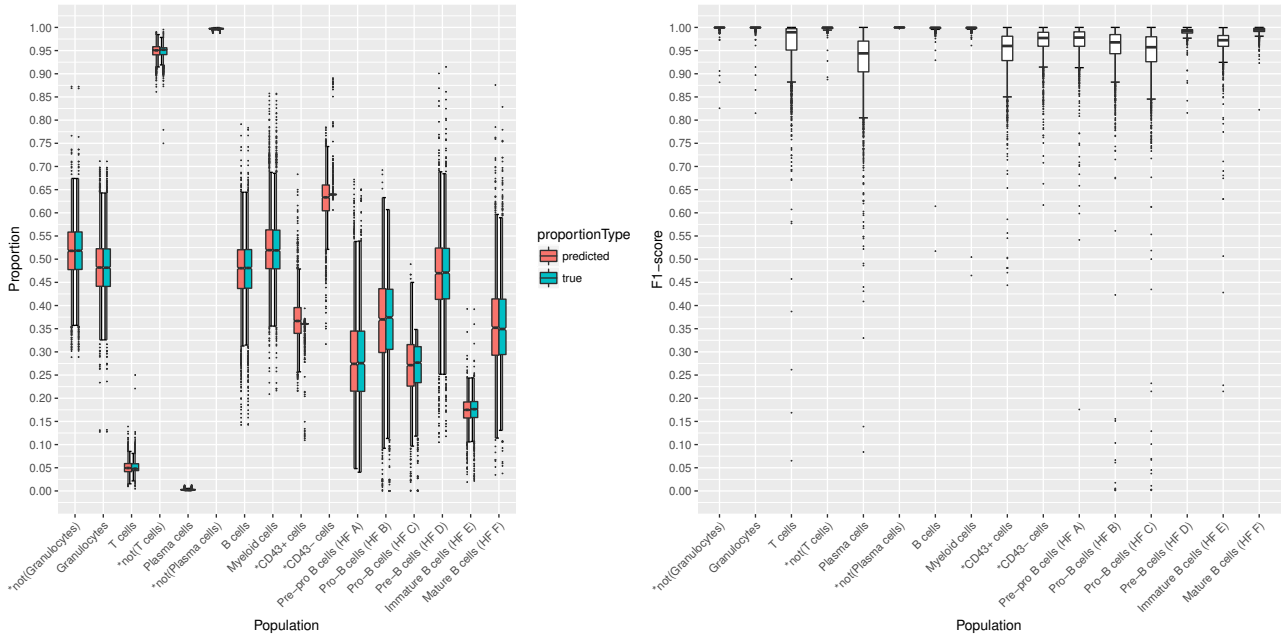


Figure 12: Results on the Mice data set for $n_p = 50$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (2665 samples). Outliers are shown as single dots. Populations HFA to HFF stand for Hardy-Fraction A to F (explained in Supplementary Table 2). Populations denoted with an asterisk are non-biological (technical) populations.

11 Full results on the FlowCAP B-cell dataset

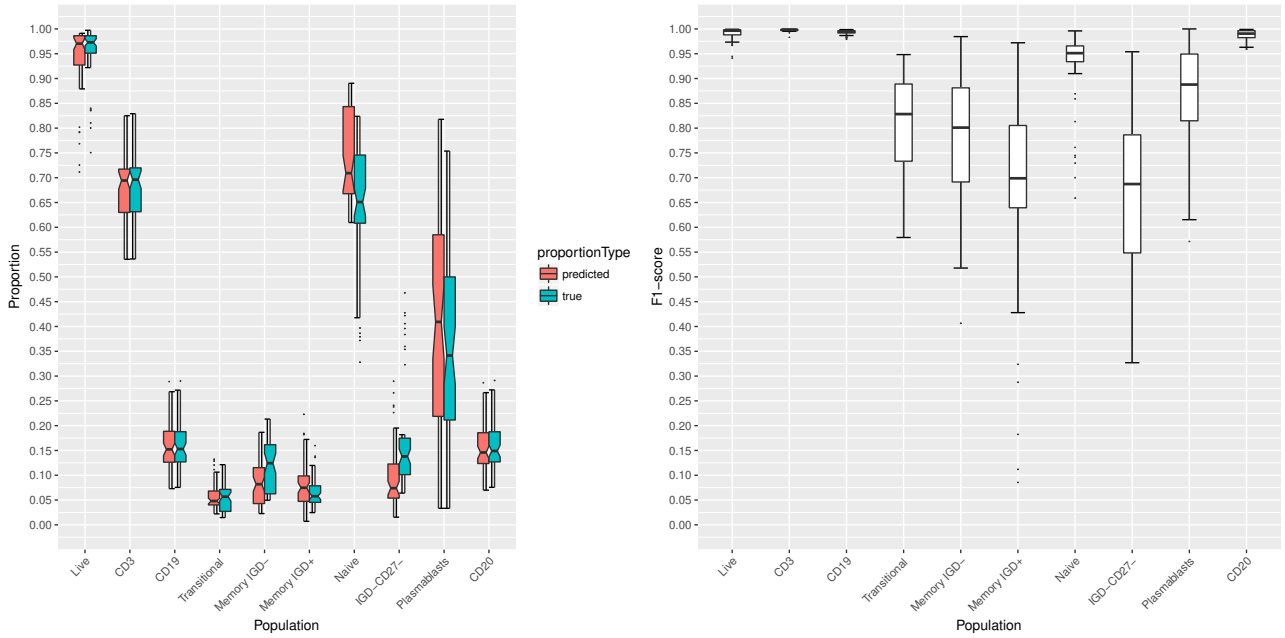


Figure 13: Results on the FlowCAP B-cell dataset for $n_p = 1$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (63 samples). Outliers are shown as single dots.

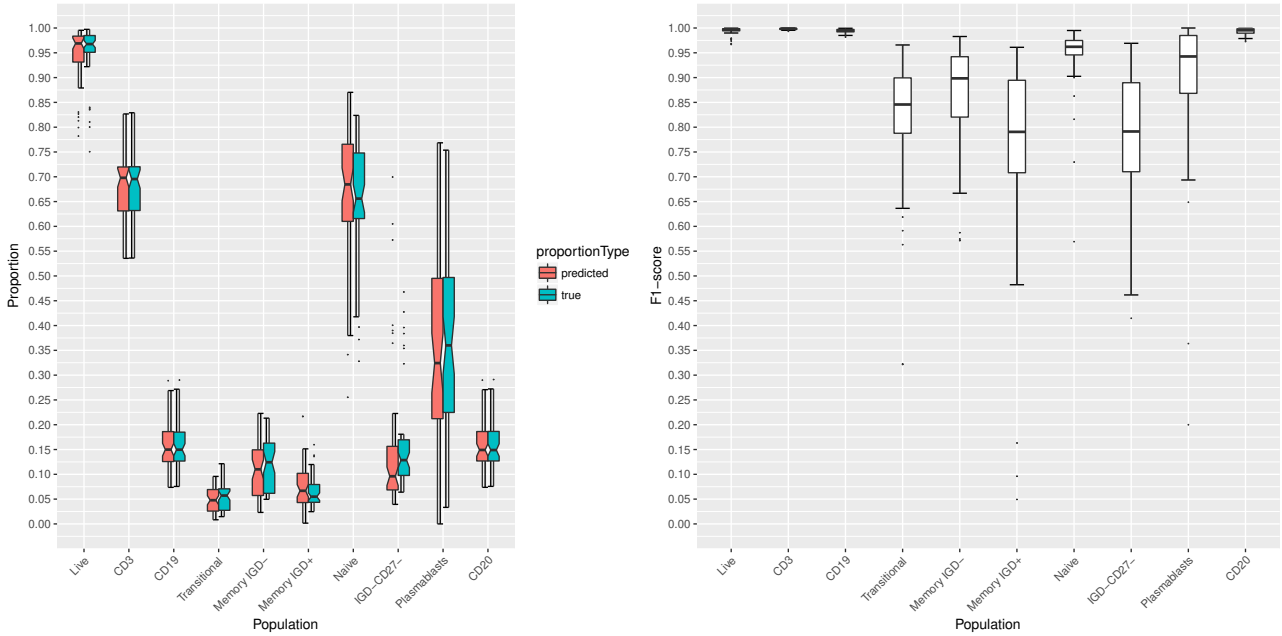


Figure 14: Results on the FlowCAP B-cell dataset for $n_p = 4$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (63 samples). Outliers are shown as single dots.

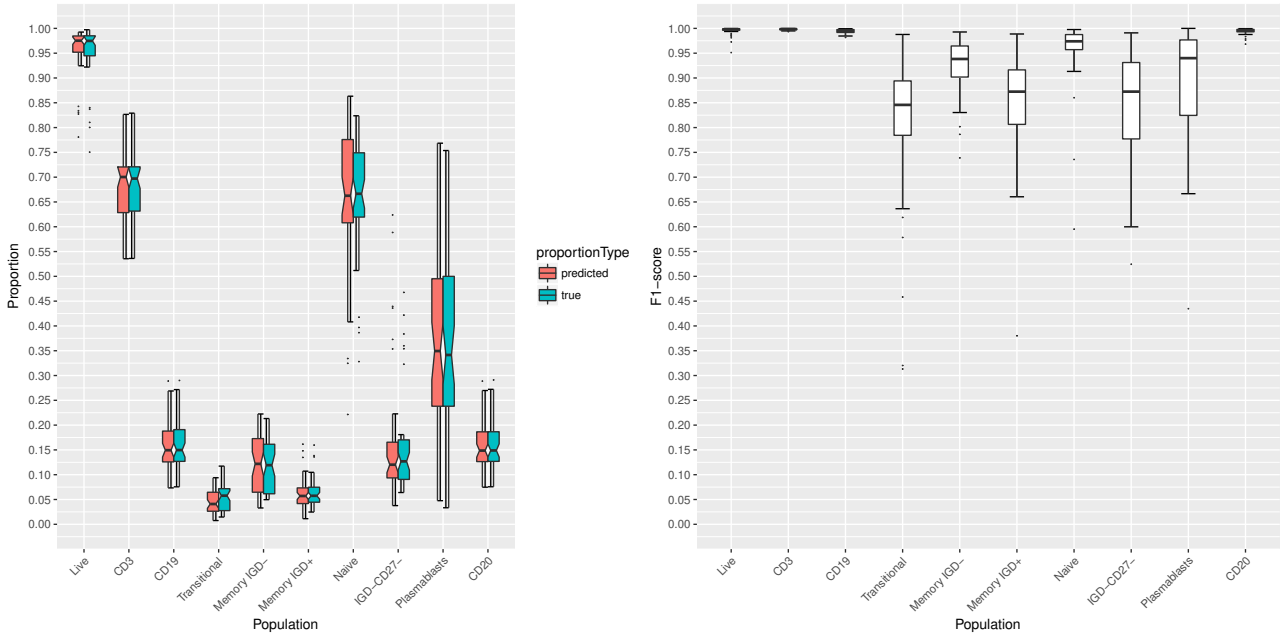


Figure 15: Results on the FlowCAP B-cell dataset for $n_p = 7$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (63 samples). Outliers are shown as single dots.

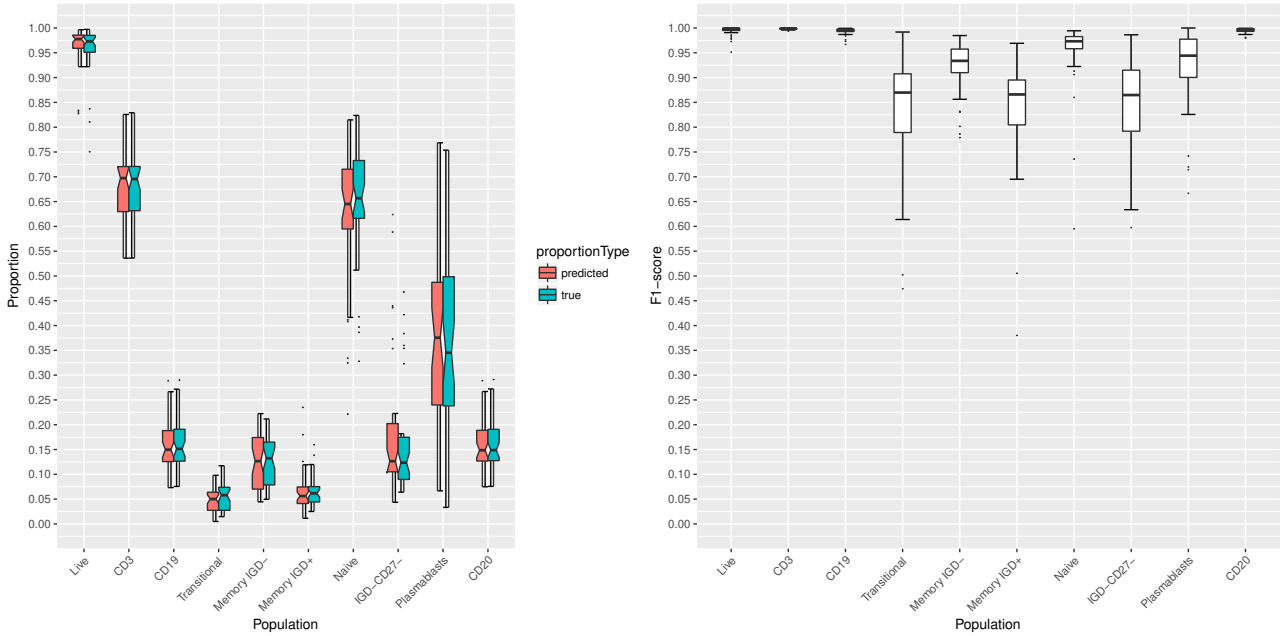


Figure 16: Results on the FlowCAP B-cell dataset for $n_p = 11$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (63 samples). Outliers are shown as single dots.

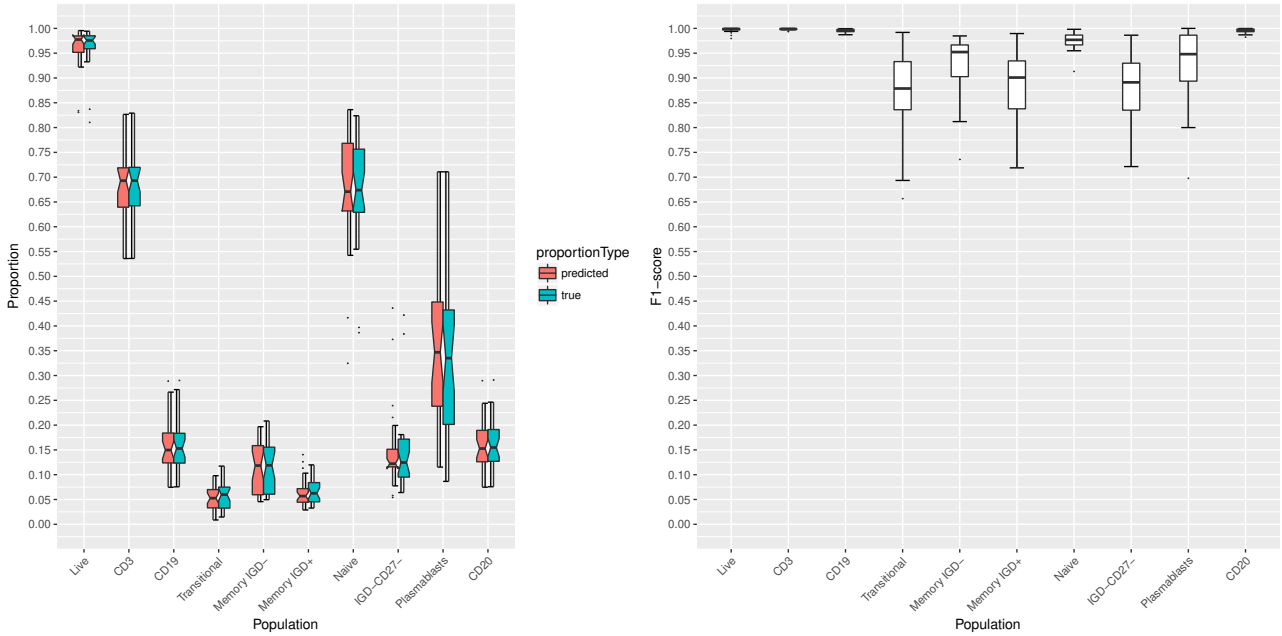


Figure 17: Results on the FlowCAP B-cell dataset for $n_p = 20$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (63 samples). Outliers are shown as single dots.

12 Full results on the FlowCAP T-cell dataset

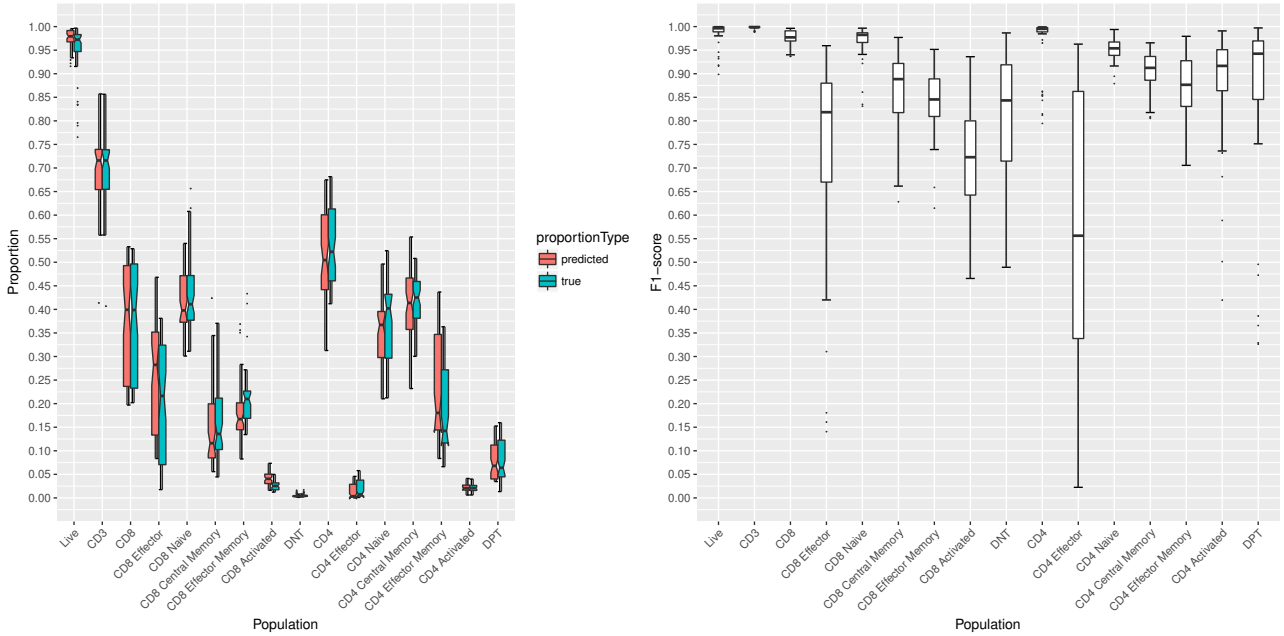


Figure 18: Results on the FlowCAP T-cell dataset for $n_p = 1$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (63 samples). Outliers are shown as single dots. Missing boxes indicate aborted computation due to failure.

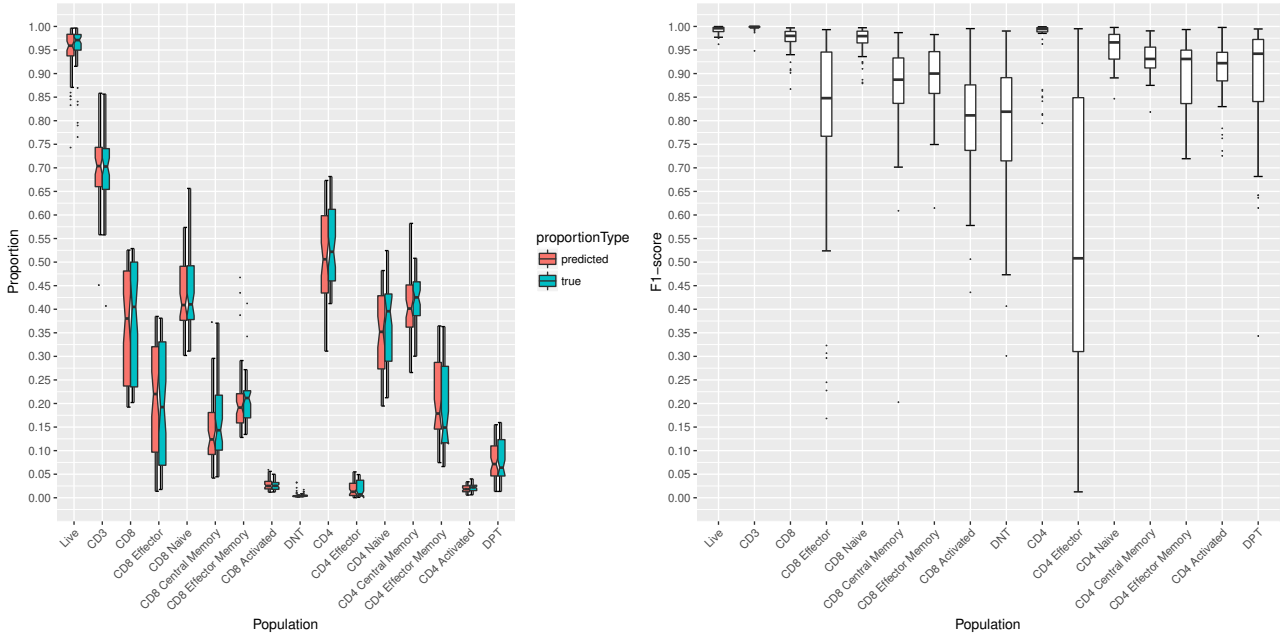


Figure 19: Results on the FlowCAP T-cell dataset for $n_p = 4$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (63 samples). Outliers are shown as single dots. Missing boxes indicate aborted computation due to failure.

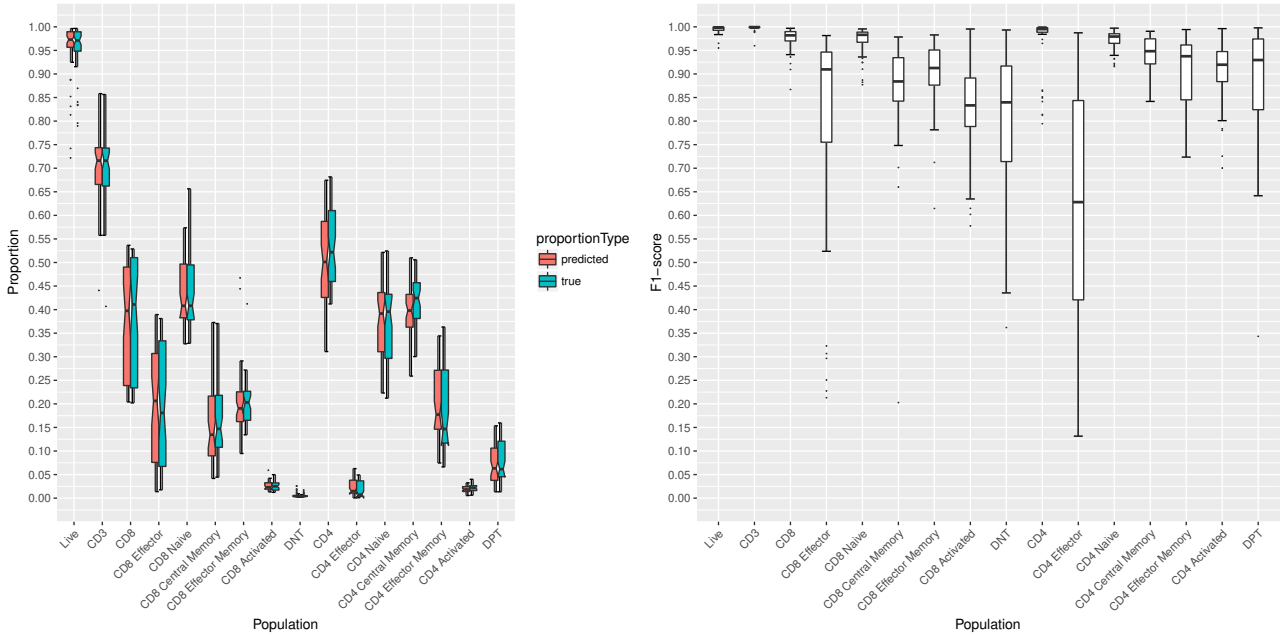


Figure 20: Results on the FlowCAP T-cell dataset for $n_p = 7$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (63 samples). Outliers are shown as single dots. Missing boxes indicate aborted computation due to failure.

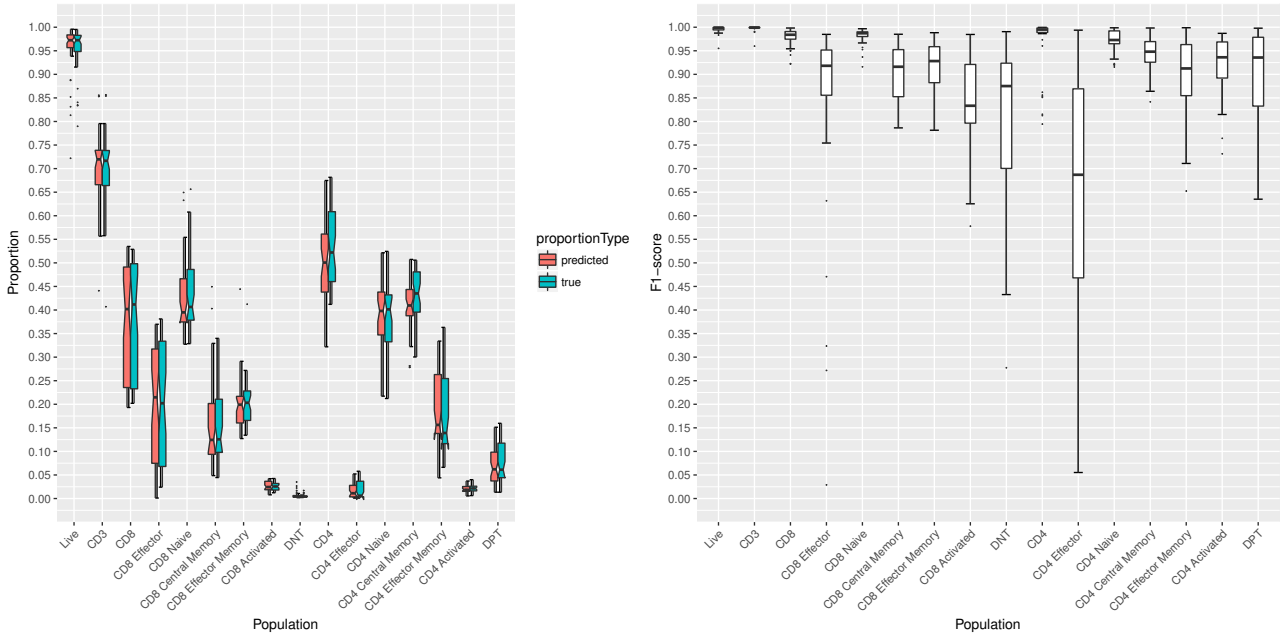


Figure 21: Results on the FlowCAP T-cell dataset for $n_p = 11$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (63 samples). Outliers are shown as single dots. Missing boxes indicate aborted computation due to failure.

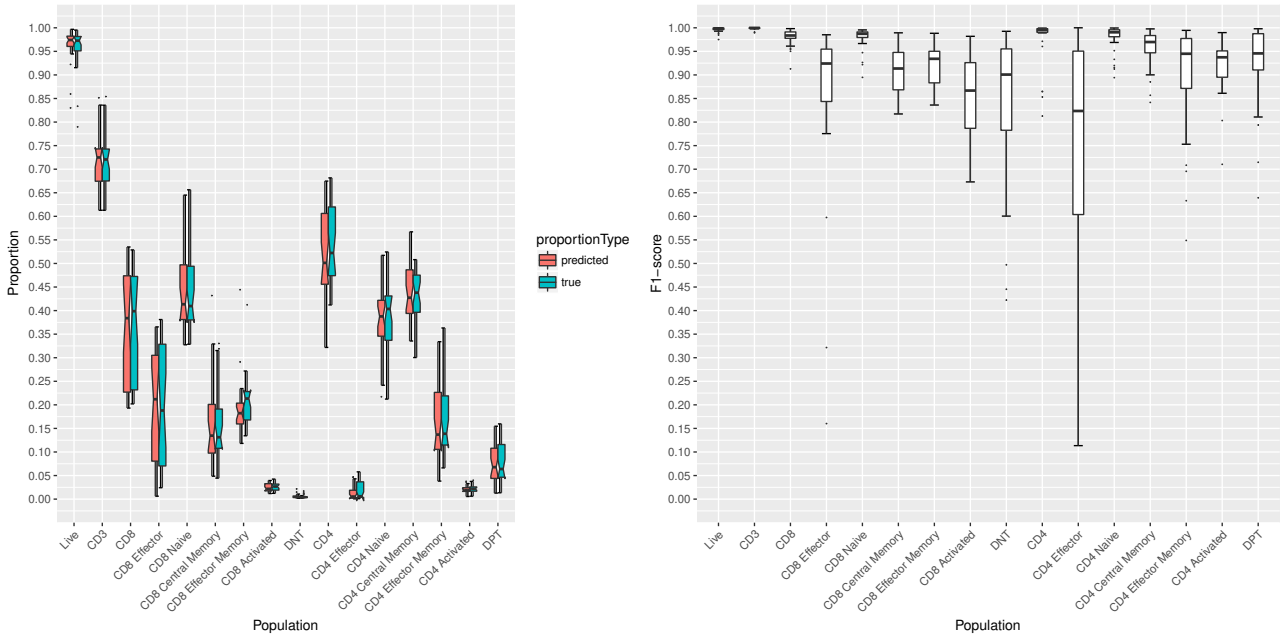


Figure 22: Results on the FlowCAP T-cell dataset for $n_p = 20$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (63 samples). Outliers are shown as single dots. Missing boxes indicate aborted computation due to failure.

13 Full results on the FlowCAP DC dataset

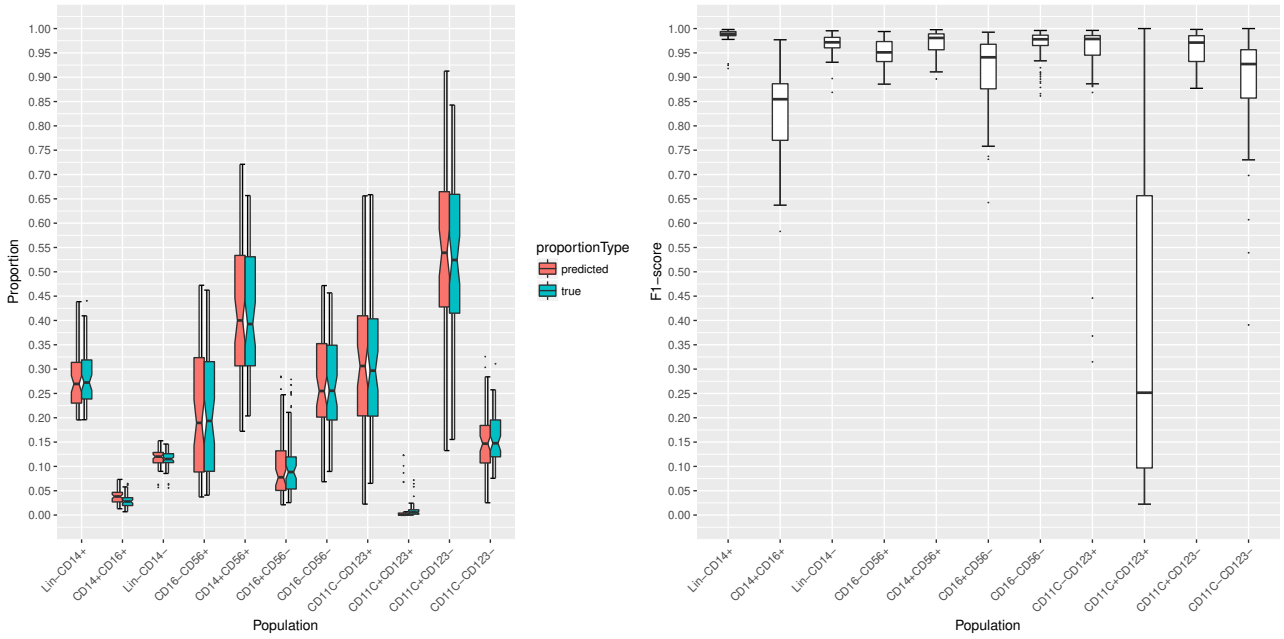


Figure 23: Results on the FlowCAP DC dataset for $n_p = 1$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (63 samples). Outliers are shown as single dots. Missing boxes indicate aborted computation due to failure.

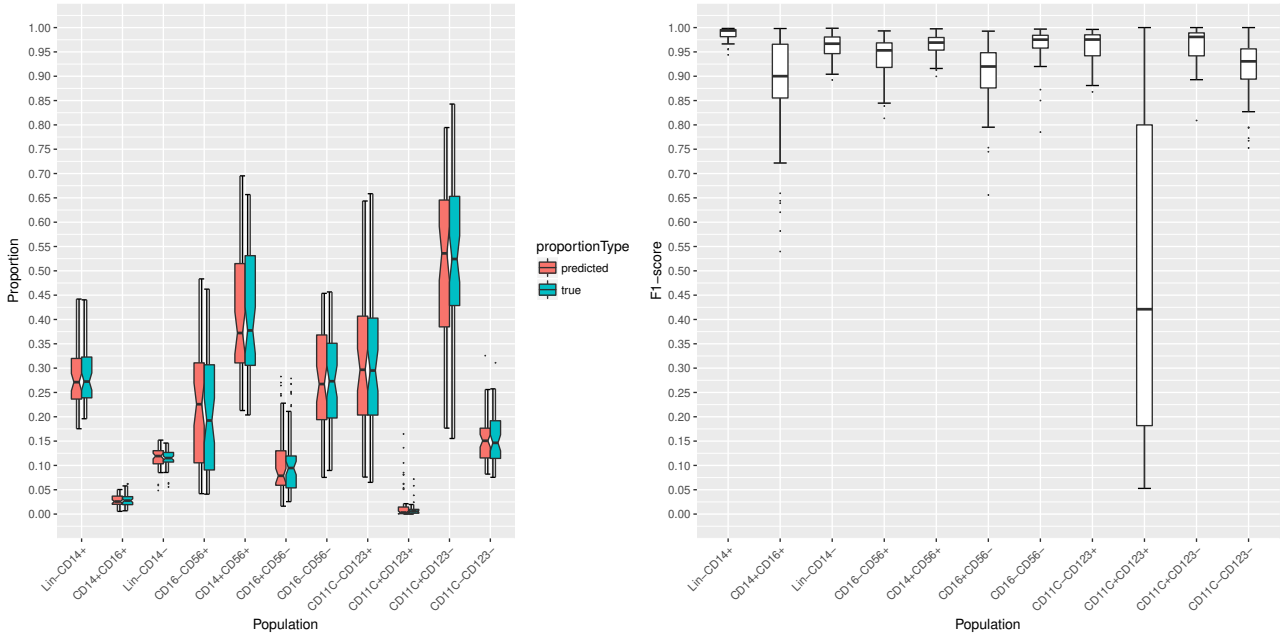


Figure 24: Results on the FlowCAP DC dataset for $n_p = 4$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (63 samples). Outliers are shown as single dots. Missing boxes indicate aborted computation due to failure.

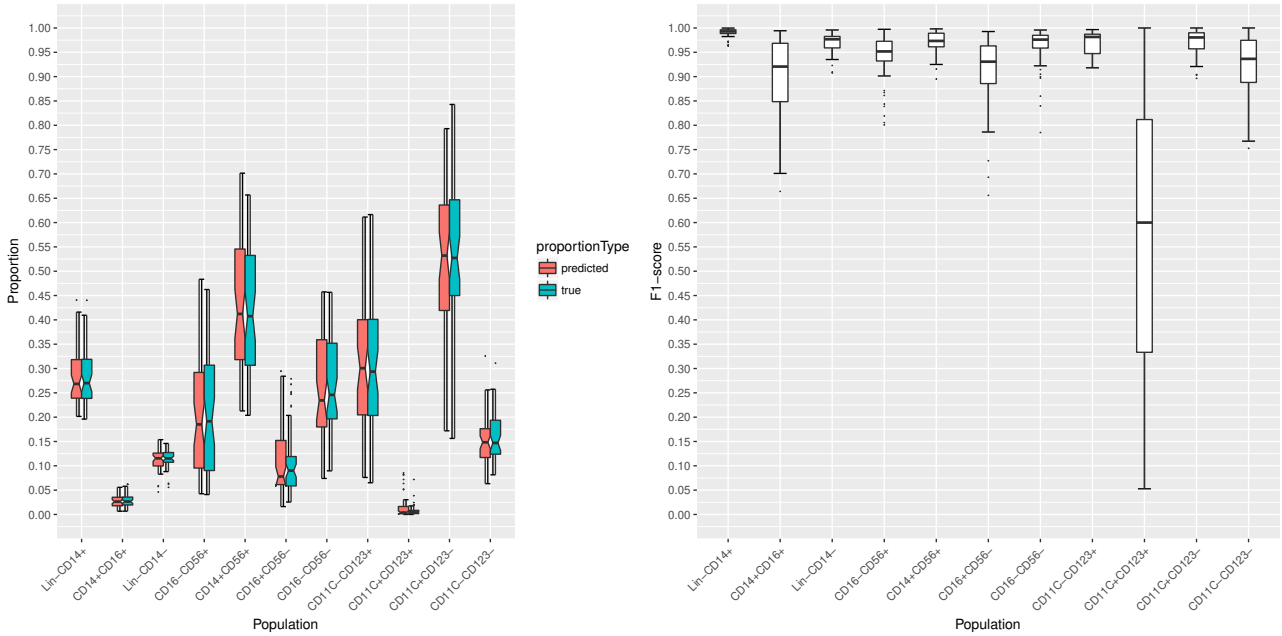


Figure 25: Results on the FlowCAP DC dataset for $n_p = 7$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (63 samples). Outliers are shown as single dots. Missing boxes indicate aborted computation due to failure.

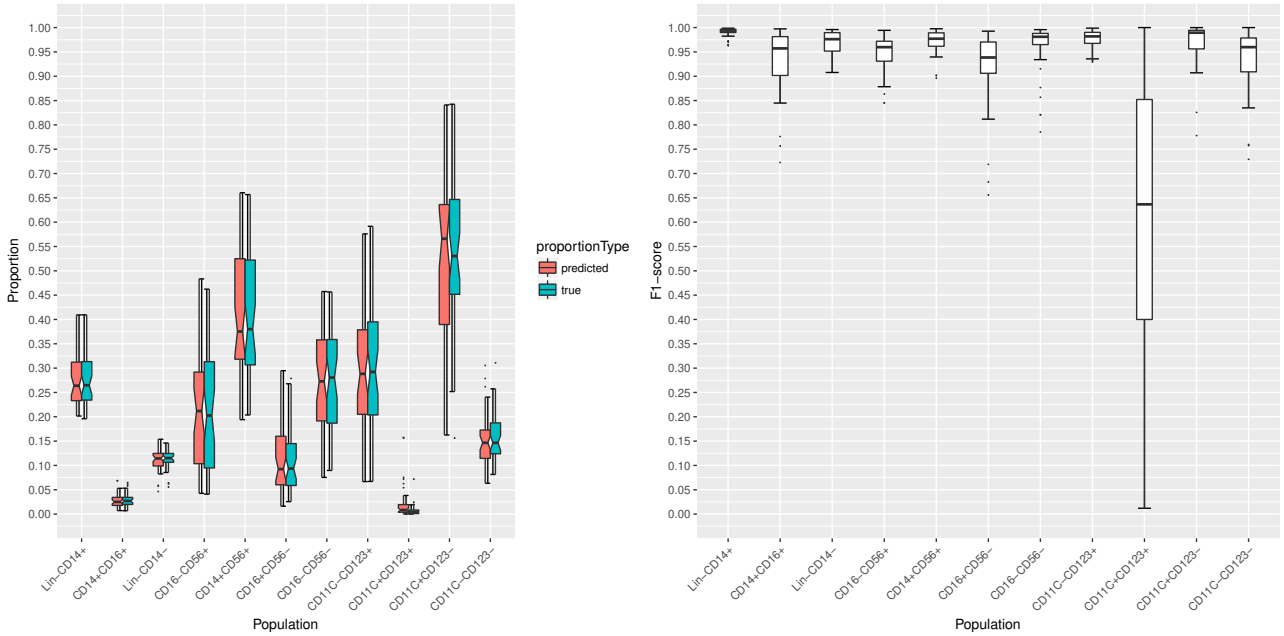


Figure 26: Results on the FlowCAP DC dataset for $n_p = 11$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (63 samples). Outliers are shown as single dots. Missing boxes indicate aborted computation due to failure.

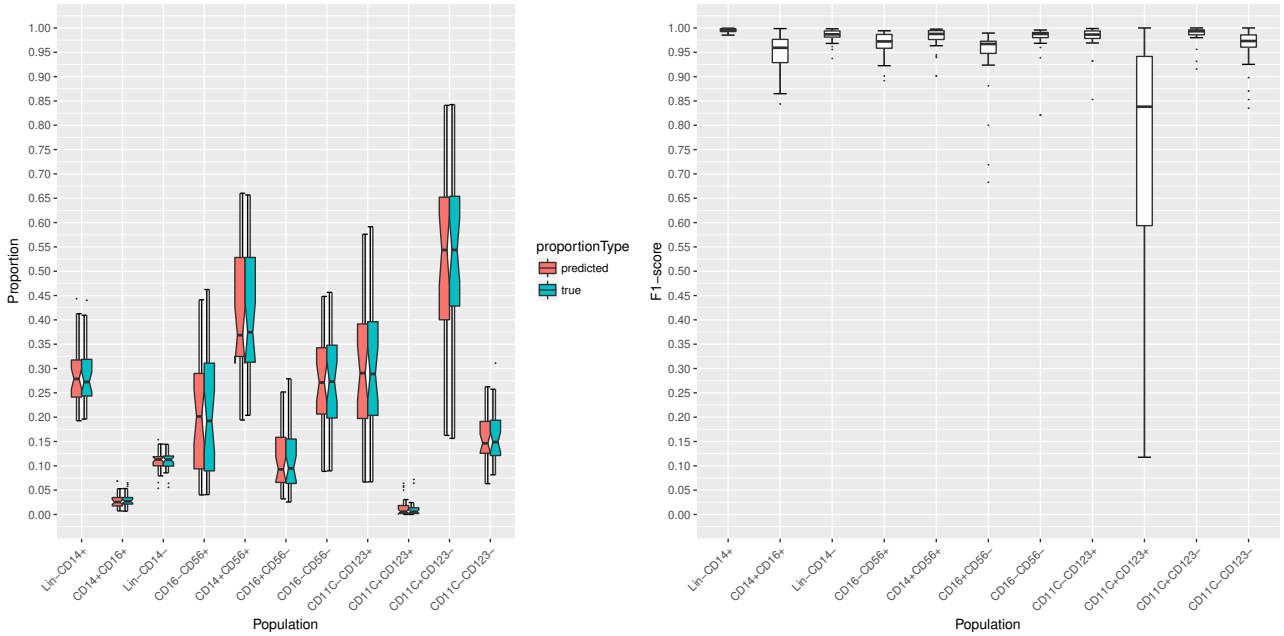


Figure 27: Results on the FlowCAP DC dataset for $n_p = 20$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (63 samples). Outliers are shown as single dots. Missing boxes indicate aborted computation due to failure.

14 Full results on the FlowCAP T-reg dataset

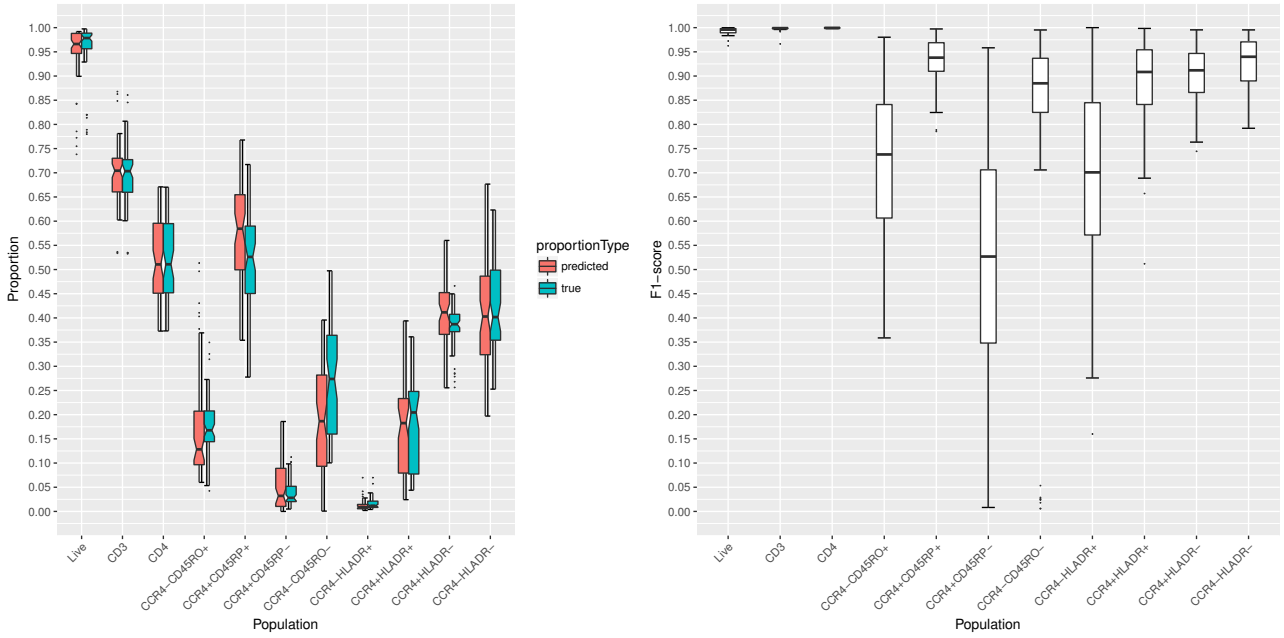


Figure 28: Results on the FlowCAP T-reg dataset for $n_p = 1$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (63 samples). Outliers are shown as single dots.

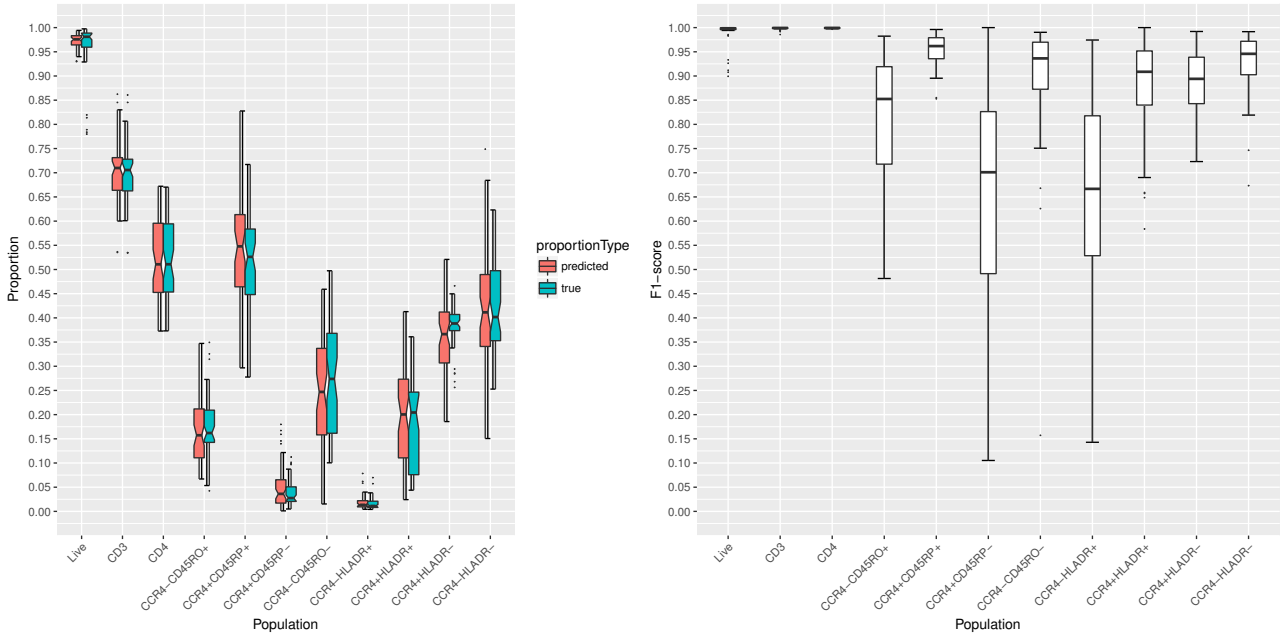


Figure 29: Results on the FlowCAP T-reg dataset for $n_p = 4$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (63 samples). Outliers are shown as single dots.

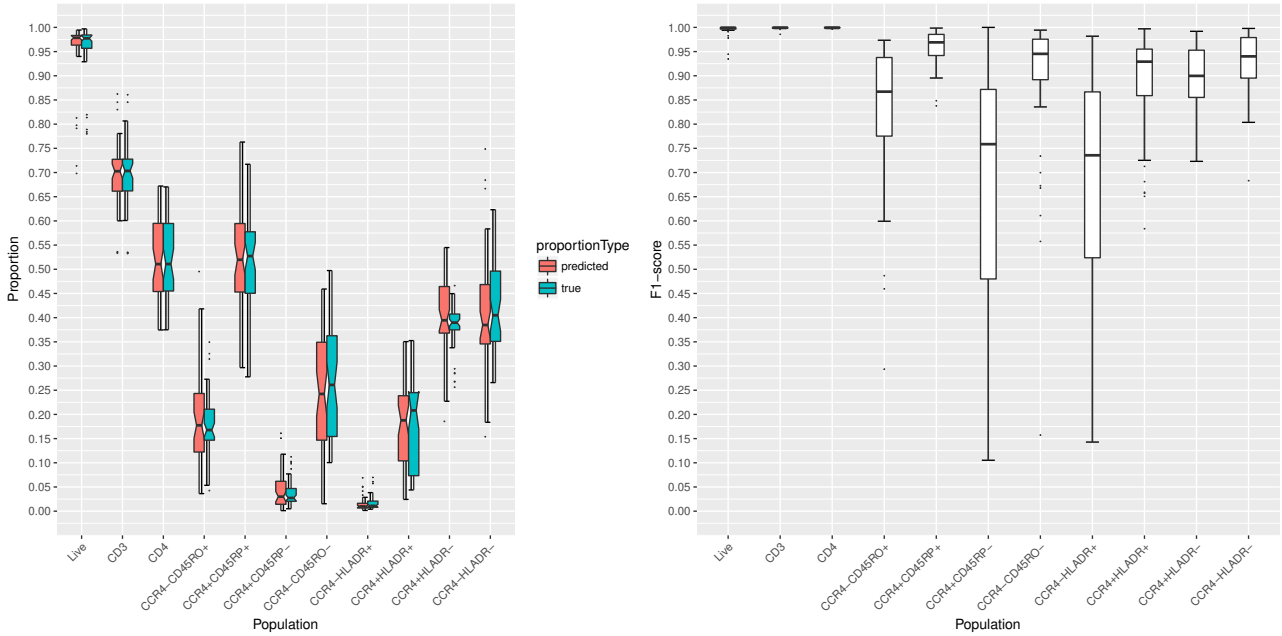


Figure 30: Results on the FlowCAP T-reg dataset for $n_p = 7$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (63 samples). Outliers are shown as single dots.

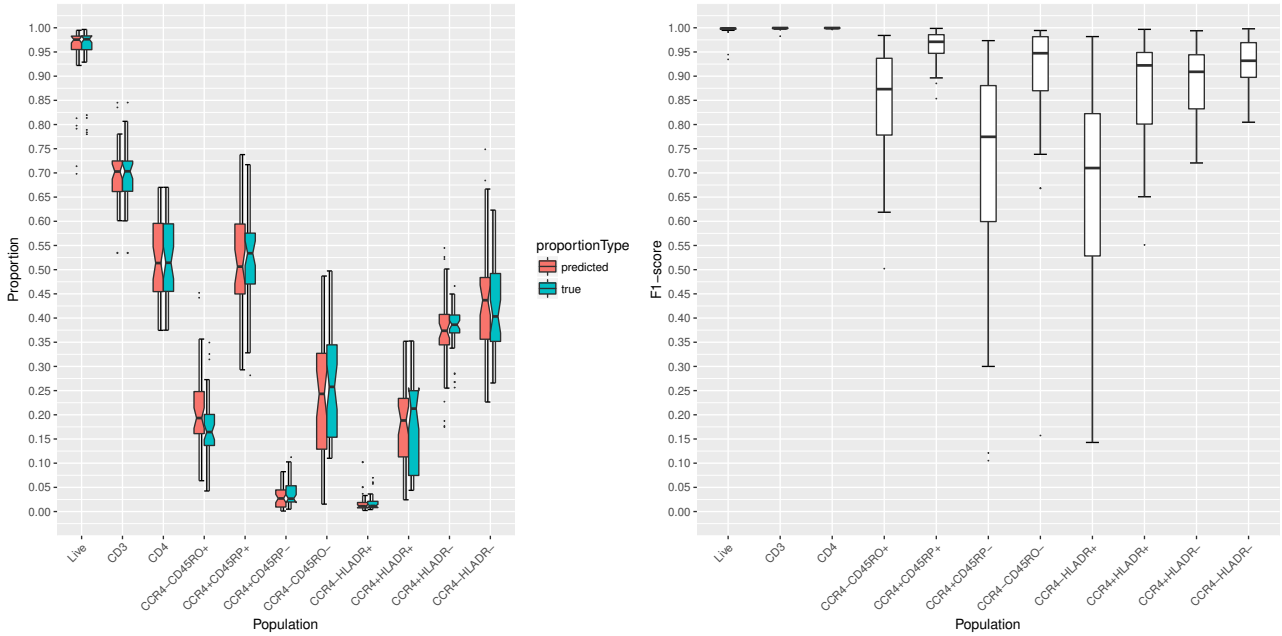


Figure 31: Results on the FlowCAP T-reg dataset for $n_p = 11$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (63 samples). Outliers are shown as single dots.

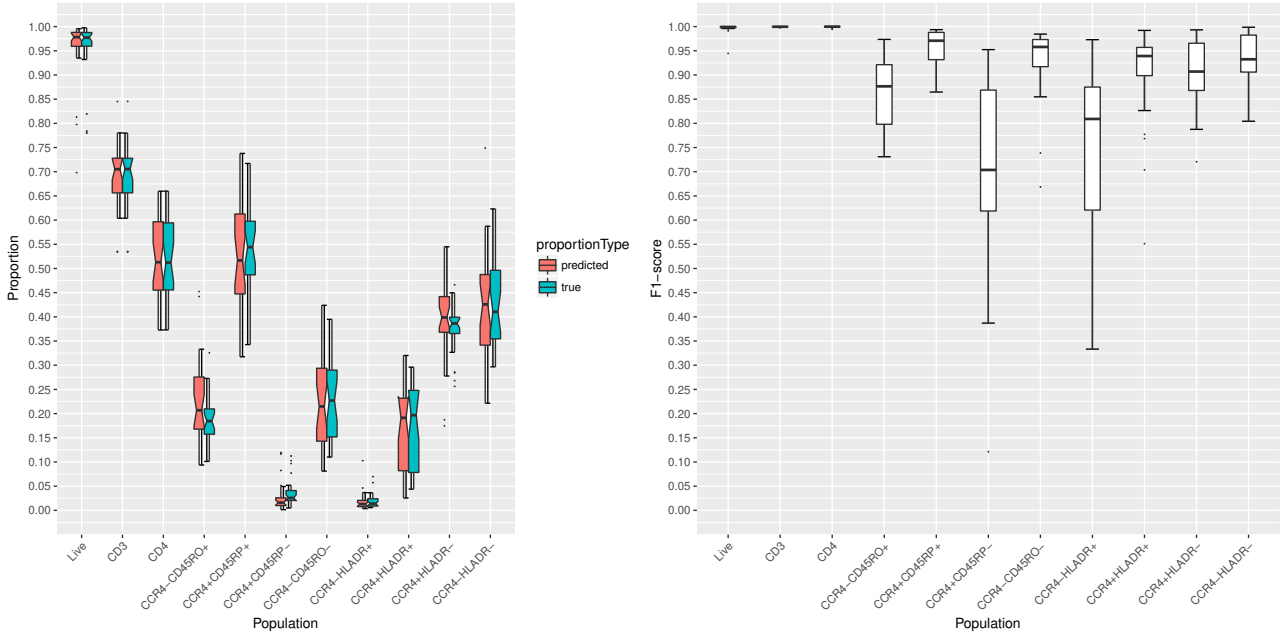
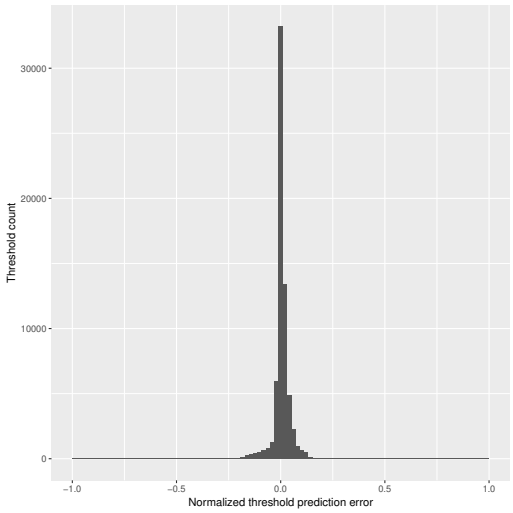
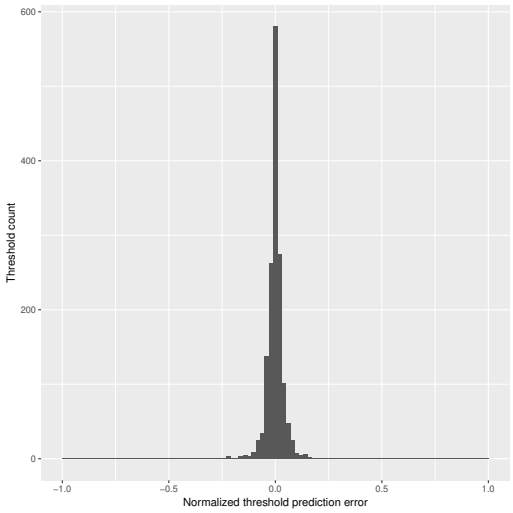


Figure 32: Results on the FlowCAP T-reg dataset for $n_p = 20$. Left: True and predicted cell frequencies for each population. Right: F_1-scores for each population (63 samples). Outliers are shown as single dots.

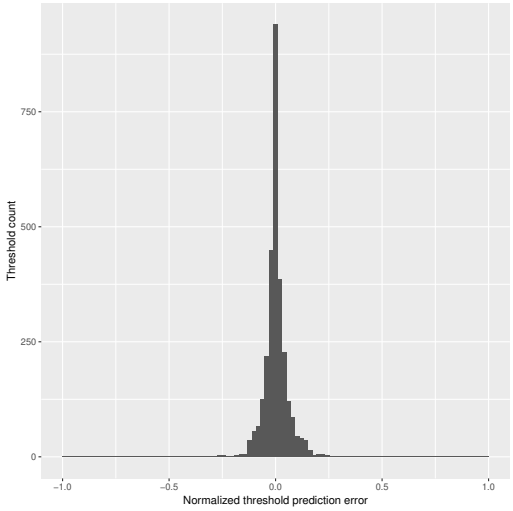
15 Distributions of predicted thresholds errors



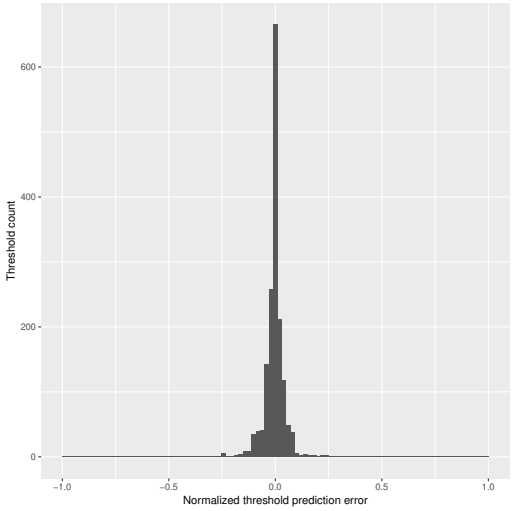
(a) Mice data for, $n_p = 1$



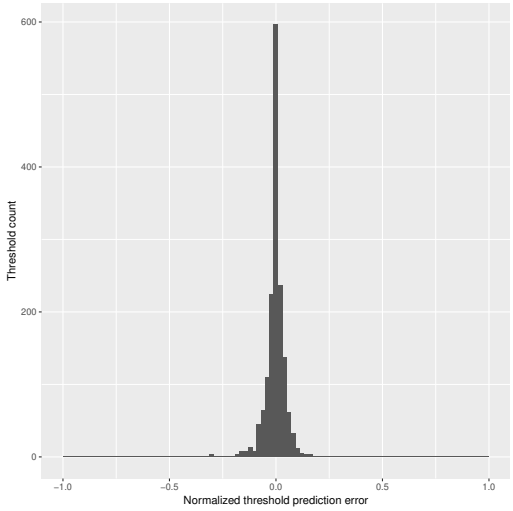
(b) FlowCAP B-cell dataset, $n_p = 7$



(c) FlowCAP T-cell dataset, $n_p = 7$



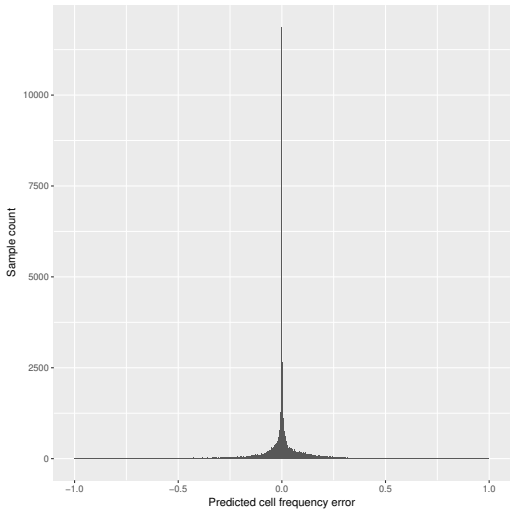
(d) FlowCAP DC dataset, $n_p = 7$



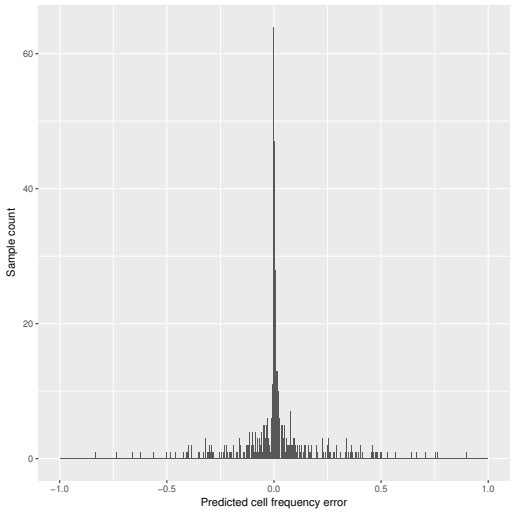
(e) FlowCAP Treg dataset, $n_p = 7$

Figure 33: Distributions of the predicted threshold error $(\hat{t_i} - t_i)/(\max d_i - \min d_i)$, normalized by the range of density d, for channels in sample i. For all predicted populations on the data sets shown in (a)–(e), respectively.

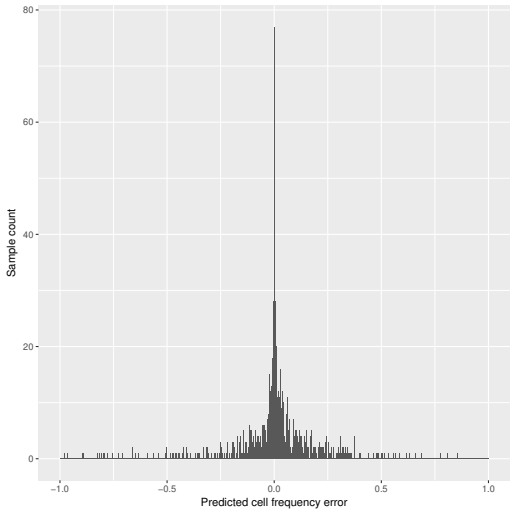
16 Distributions of predicted cell frequency errors



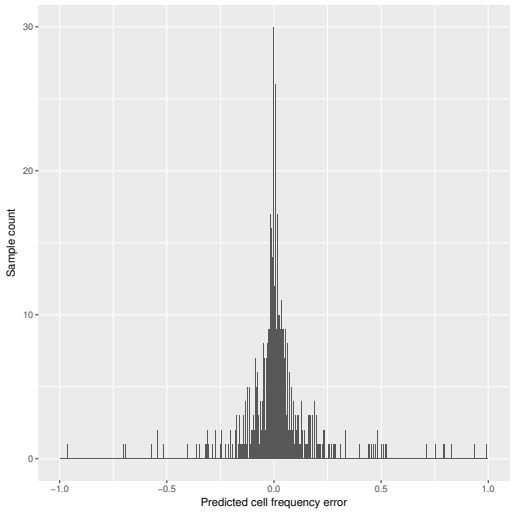
(a) Mice data for, $n_p = 1$



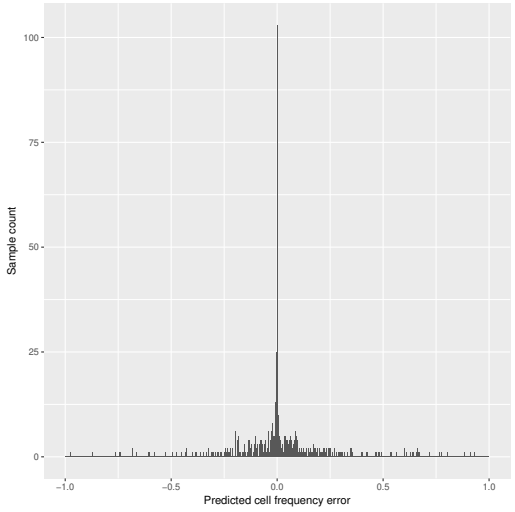
(b) FlowCAP B-cell dataset, $n_p = 7$



(c) FlowCAP T-cell dataset, $n_p = 7$



(d) FlowCAP DC dataset, $n_p = 7$



(e) FlowCAP Treg dataset, $n_p = 7$

Figure 34: Distributions of the predicted cell frequency error $(f_i - \hat{f}_i)/(\text{median } f_i)$ for sample i, for all predicted populations on the data sets shown in (a)–(e), respectively.

17 Explanation of HFA – HFF populations

Throughout the manuscript, with HFA – HFF populations, we refer to different developmental stages of B-cells in bone marrow. While the division of B cell precursors into Hardy fractions does not fully capture the current discussions on the complexity of B cell development, it is a concept that many immunologists are familiar with and can easily relate to.

Table 3: Alternative names for HFA–HFF populations

Population	Alternative name	Hardy fraction
HF A	Pre-pro-B cells	(Hardy fraction A)
HF B	Pro-B cells	(Hardy fraction B)
HF C	Pro-B cells	(Hardy fraction C)
HF D	Pre-B cells	(Hardy fraction D)
HF E	Immature B cells	(Hardy fraction E)
HF F	Mature B cells	(Hardy fraction F)