Differences between CellScanner 1.0 and 2.0

Functions
CellScanner 1.0 has several functions that CellScanner 2.0 does not implement. Their omission is justified in the brackets. These are:

- Species labelling through unsupervised classification (we saw that the accuracy of this procedure is significantly lower than supervised classification, see supplementary Figure 1 in the evaluation)
- Two additional classifiers, namely random forest and logistic regression (in the evaluation, these did not perform better than neural networks)
- Database of flow cytometry files (CellScanner 2.0 allows easy loading of sets of FC files per species and repeated application of a trained model to co-culture files)
- Sub-selection of channels used for classification (we saw that the channels that are important for the prediction differ between species combinations, see Figure 6 in https://www.nature.com/articles/s43705-022-00123-6)
- Tool analysis (this is the computation of the confusion matrix, which is now done automatically whenever a classifier is trained)
- Flow cytometer definition (this has become obsolete with the other changes)

CellScanner 2.0 also has functions absent in CellScanner 1.0, namely the support for dead/live and DNA staining, which allows removal of dead cells and events without a DNA signal given user-provided thresholds. In addition, it computes the heterogeneity of events after gating.

Preprocessing
Both versions drop the ‘Time’ column from flow cytometry files. CellScanner 2.0 in addition applies the arcsinh transformation (with division by a constant factor set to 150) followed by a Z-transformation.

Gating
CellScanner 1.0 offers three gating routines, referred to as line gating for Accuri, line gating for CytoFlex and machine gating. The line gating routines hard-code values that define lines to separate debris from cells and are specific to Accuri and CytoFLEX flow cytometers, respectively. In the evaluation, they performed less well than machine gating. Machine gating trains the selected classifier on the blanks and the non-blanks to distinguish cells from debris. The classifier is run 6 times on 1000 events from each category and uses majority voting of 70%. The problem with this approach is that some events in the non-blanks represent debris, too. Thus, in CellScanner 2.0, unsupervised classification with UMAP is used to cluster debris separately from cells. If staining information is provided, it is taken into account to separate live from dead cells and debris.

Assessment of uncertainty

In CellScanner 1.0, a species label is determined by running a classifier (e.g. random forest) several times on different data sub-sets (which can overlap, i.e. bootstrapping) and aggregating the results through a majority vote. Uncertainty is assessed through the number of runs that do not agree. We have shown in the evaluation that prediction accuracy can be improved by increasing the percentage of runs that need to agree on a classification (e.g. from 7 to 9 out of 10). In CellScanner 2.0, uncertainty is reported as Shannon entropy, which is computed based on the probability vector across labels (species, blank) that the classifier (the neural network) returns for each event.