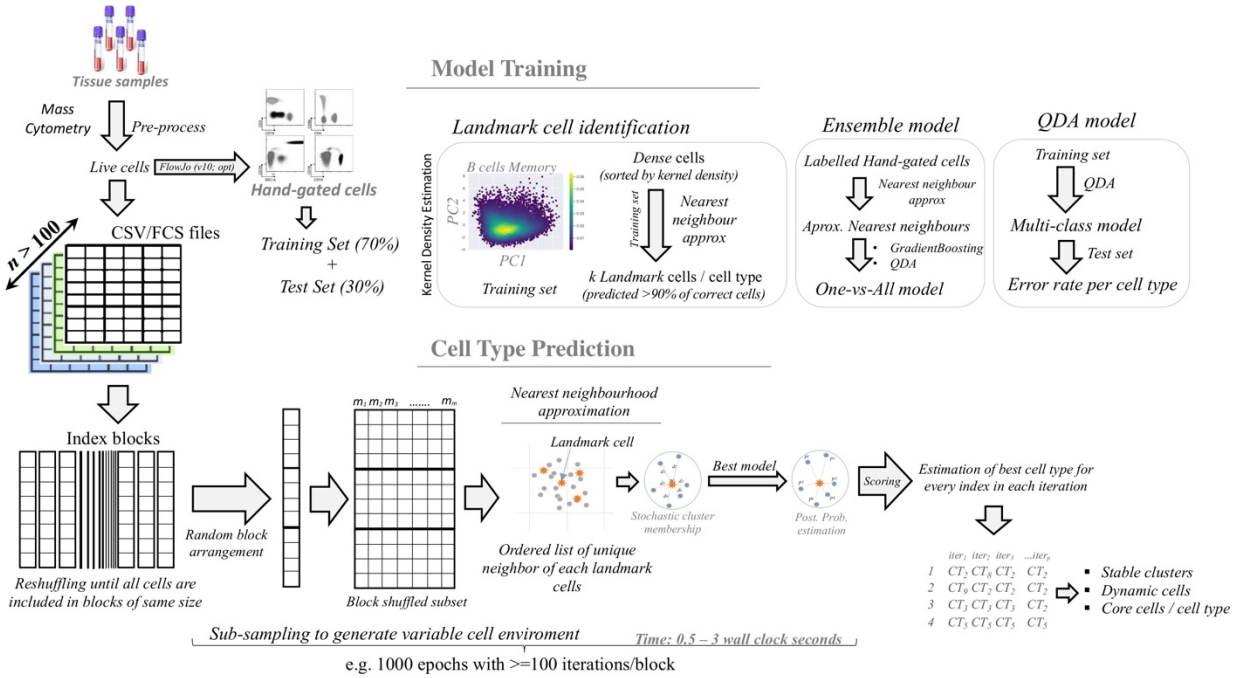
CyTOF-annotator: Workflow



Data set: A subset of manually hand-gated cell types were used for model training. Remaining labeled files are used as an independent dataset for algorithm testing and validation.

Model training*:* The algorithm begins with the following multi-step procedure to calculate local and non-linear boundaries for each cell type using the hand-gated cells:

- *Splitting*: The hand-gated cells are pooled and then split into- training (70%) and test set (30%), wherein the cell type label proportion remains the same in both the sets.
- Quadratic Discriminant Analysis: Subsequently, a multi-class QDA model is built using labels from training set, of which the accuracy per cell type is measured in testing dataset.
- *Landmark cells identification:* In the training set, the cell-cell closeness per cell type are identified by decomposing the high-dimensional data points into two Principal Components (PCs) followed by estimating the kernel densities of each cell. Using this, the total number of landmark cells per cell type is determined as the set of non-similar cells from the training set with following properties:
    1. Landmarks cells are picked from different Kernel densities region, with preference from higher density to lower density region.
    2. Landmark cells from same density quantile range are distant from each other, i.e. no two landmark cells are close to each other, rather few landmark cells spread across entire PCs scatter plot.
    3. Landmark cells can re-predict at least 20% of true positive cells by nearest neighborhood approximation method using the test set.
    4. No two landmark cells of a given cell type predict more than and equal to 90% of common true positive cells, i.e. landmark cell(s) with redundant prediction result will not be considered.

5. Minimum number of landmark cells that can find >99% of true positive cells on test set are selected.

- *OneVsAll Cell type prediction model*: For each cell type, the selected landmark cells are used to predict nearest neighbors from the whole set of hand-gated cells. The approximated neighbors are then re-labelled, such that true positive cells are labelled as '1' whereas rest of the nearest neighbors of landmarks are labelled as '0'. This new dataset with binary labels is then split into training (70%) and testing (30%) dataset. The training set is then used to build GradientBoosting and QDA models. The best model with highest True-Positive rate is used for the given cell-type.

Cell type prediction:

- *Index Blocking*: Each sample of independent live cells dataset is split into a defined number of blocks of equal numbers of randomly selected unique cell.
- *Bootstrapping*: In each iteration, randomly selected blocks, one from each sample, are shuffled to create a unique sub-sample that undergoes cell type identification steps. Next, the approximated nearest neighbor cells for each landmark are determined. In each iteration, the approximated nearest neighbor cells are score by following function:

$$Score_c^{ct} = \left(w_i * \mathrm{d}_j\right) + p$$

Where, for cell *c*, *w* is the normalized kernel density of closest landmark cell *i*, and d is the approximate distance to that landmark cell. In a given iteration, a cell *c* is assigned to cell type *ct* with highest value of scoring function. The total cell-type membership score after all iteration for every cell-type in which a cell can occur is then calculated. The credibility of the prediction for each cell is determined by evaluating the variance in the cell types found to be associated with it across all epochs.

Cluster stability:

- The cluster stability and the core cells associated with stable clusters are predicted by enumerating the variations associated with its cells. For downstream biological analysis, only the stable cells enriched within a given cell type may be used for reliable estimation of the differential behavior in their functional markers.