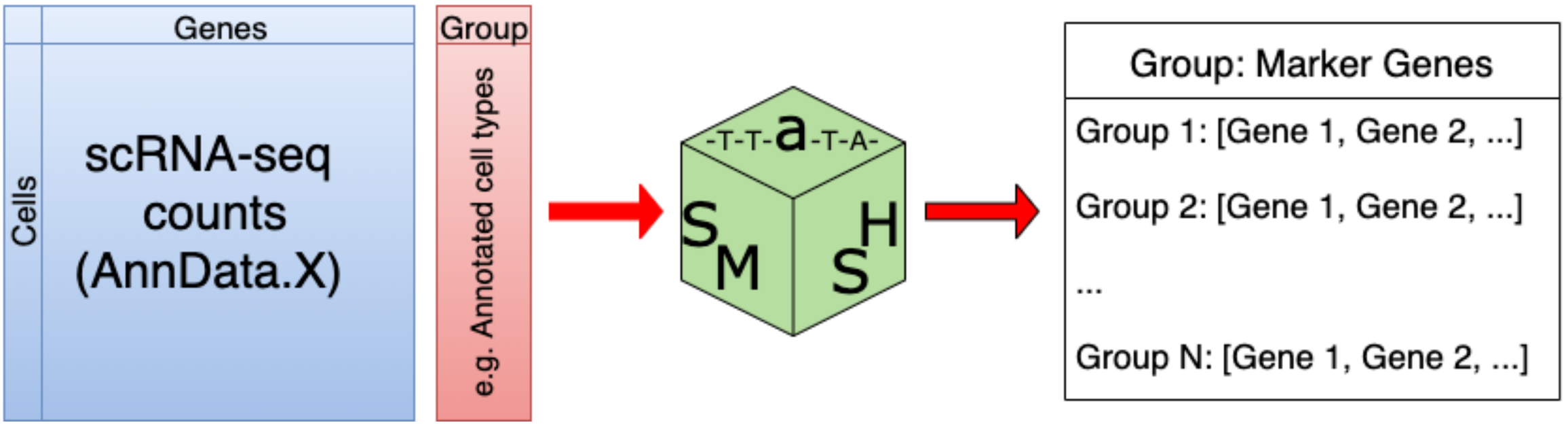
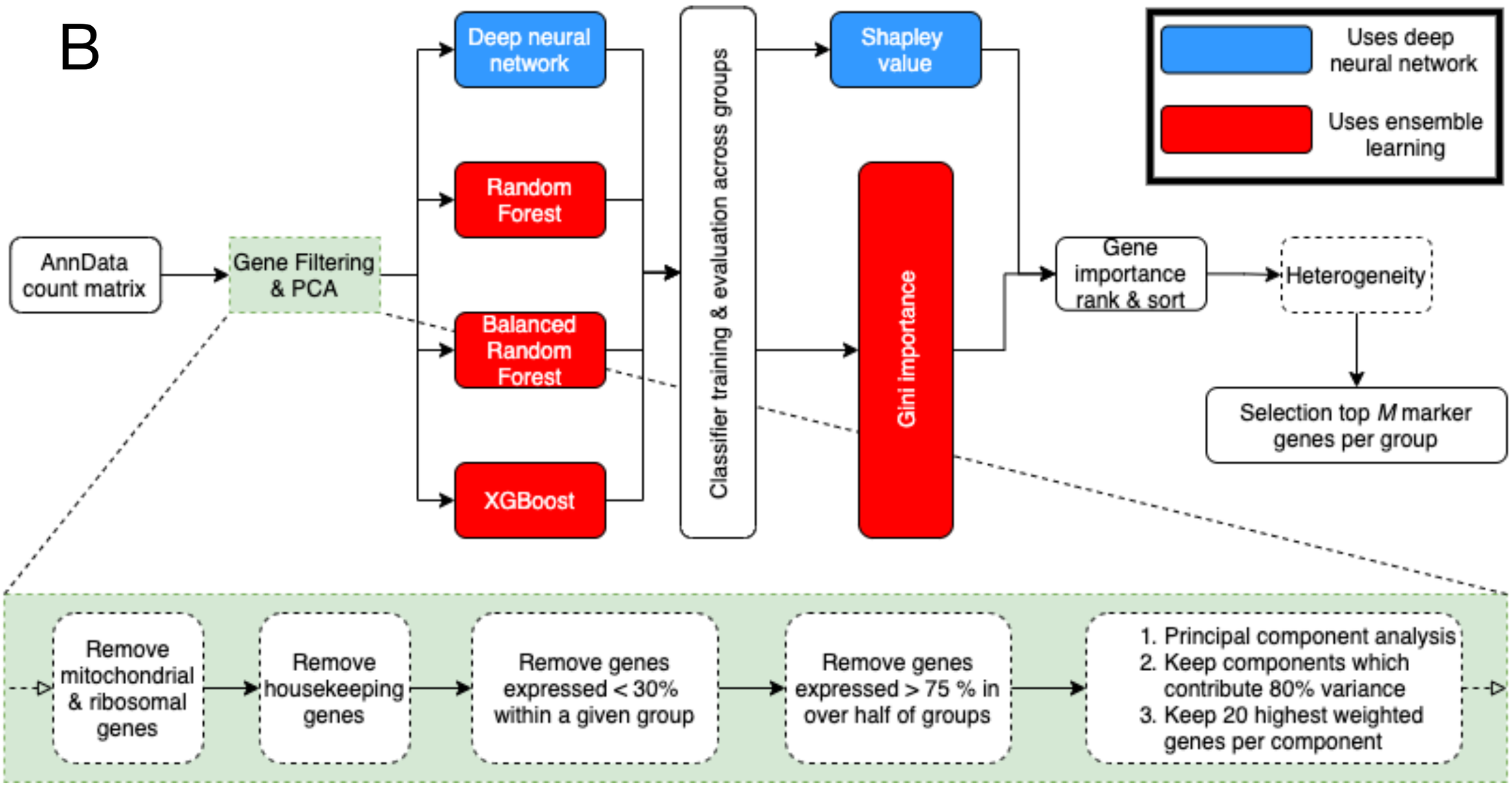
A


Genes
Cells
scRNA-seq counts (AnnData.X)
Group
e.g. Annotated cell types
-T-T-a-T-A-
S M
S H
Group: Marker Genes
Group 1: [Gene 1, Gene 2, ...]
Group 2: [Gene 1, Gene 2, ...]
...
Group N: [Gene 1, Gene 2, ...]

B


Deep neural network
Random Forest
Balanced Random Forest
XGBoost
Classifier training & evaluation across groups
Shapley value
Gini importance
Uses deep neural network
Uses ensemble learning
AnnData count matrix
Gene Filtering & PCA
Gene importance rank & sort
Heterogeneity
Selection top M marker genes per group
Remove mitochondrial & ribosomal genes
Remove housekeeping genes
Remove genes expressed < 30% within a given group
Remove genes expressed > 75 % in over half of groups
1. Principal component analysis
2. Keep components which contribute 80% variance
3. Keep 20 highest weighted genes per component