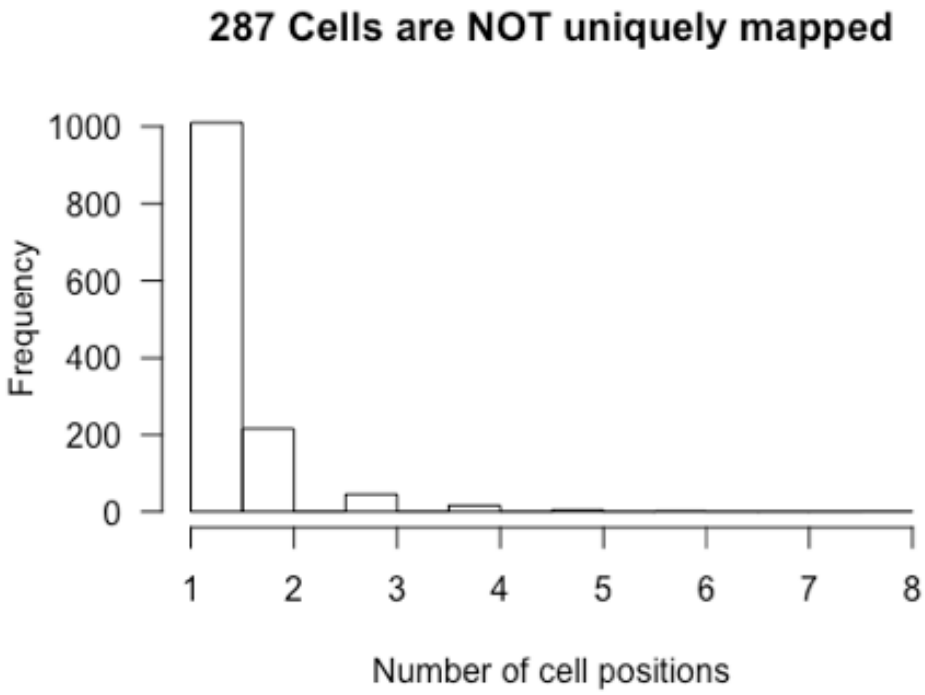
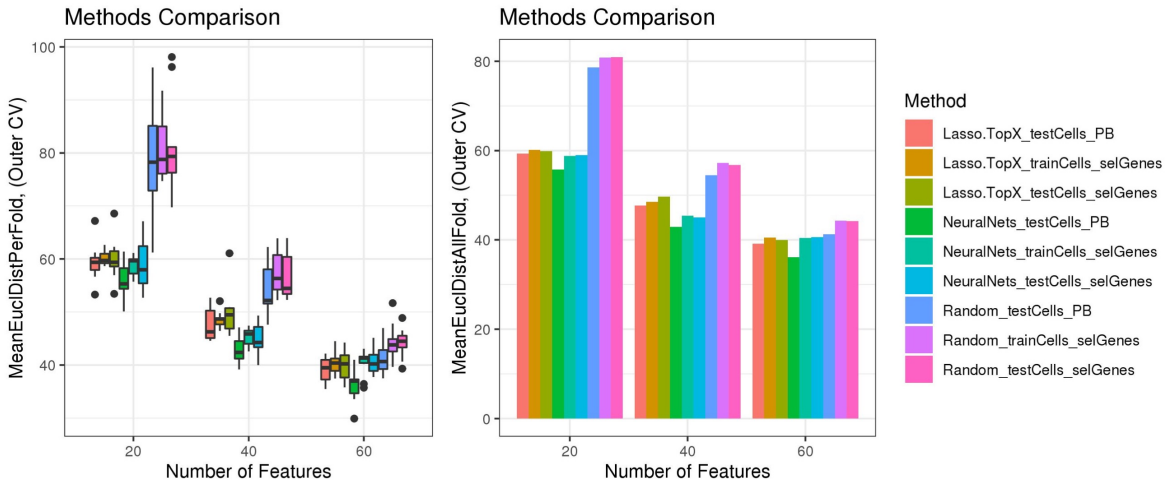
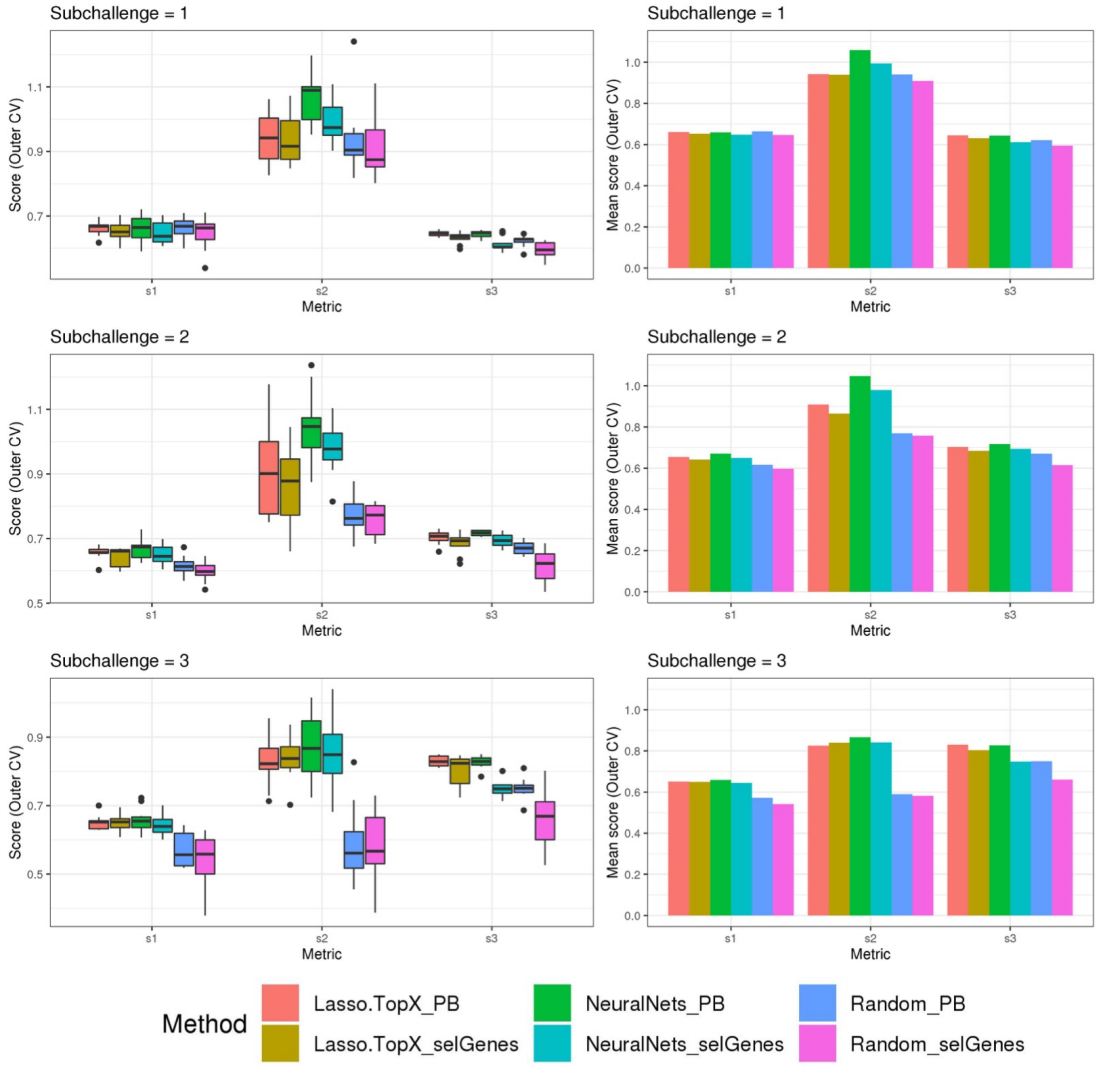
Supplementary Material of Machine Learning Approaches Identify Genes Containing Spatial Information From Single-Cell Transcriptomics Data



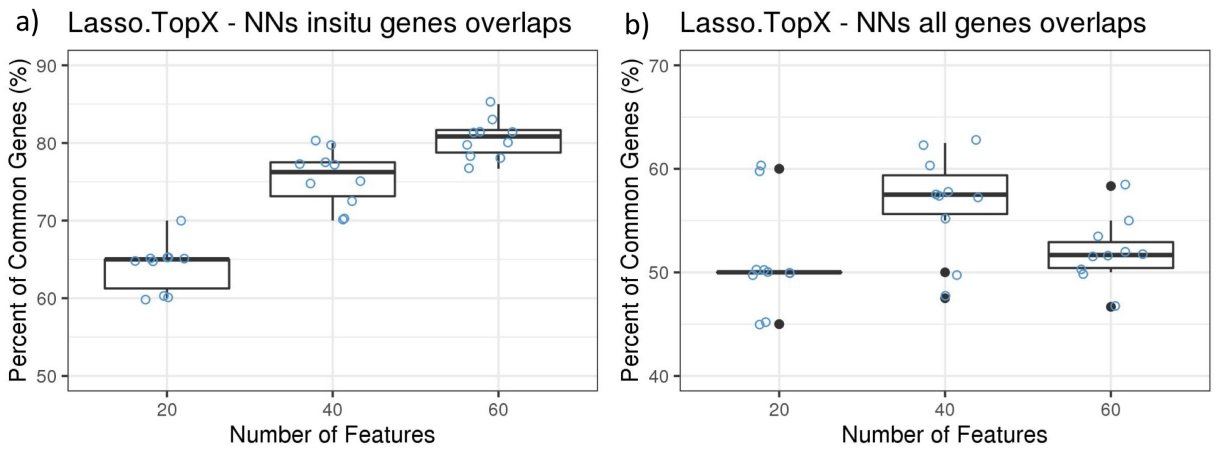
Supplementary Figure S1. DistMap (using all 84 inSitu genes) reports 287 cells that share their max-MCC position with another cell



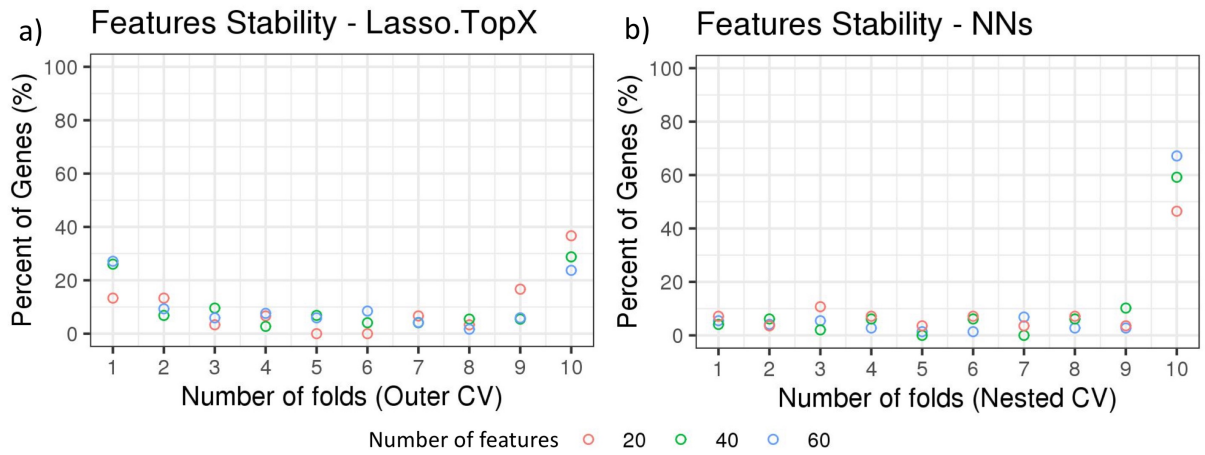
Supplementary Figure S2. Comparison of different methods using Blind metric. Lasso.TopX (Lasso.TopX_testCells_selGenes) performed better on selecting 60 features, sub-challenge 1, whereas NeuralNets (Light Blue - NeuralNets_testCells_selGenes) performed better in selecting 40 and 20 genes, sub-challenges 2 and 3. Training and testing errors present small differences which point to there is no overfitting taking place in either Lasso.TopX (Lasso.TopX_trainCells_selGenes and Lasso.TopX_testCells_selGenes) or NeuralNets (NeuralNets_trainsCells_selGenes and NeuralNets_testCells_selGenes). Both methods performed better than Random (Random_trainCells_selGenes and Random_testCells_selGenes). Using the provided binarized expression data, _PB extension, leaded to overestimation of performance accuracy, Lasso.TopX_testCells_PB, NeuralNets_testCells_PB, Random_testCells_PB. Xaxis, different number of features (e.g. 60/40/20 for subchallenge #1/#2/#3 respectively). Yaxis, left figure, MeanEuclDistPerFold, Mean Euclidean Distance Per outer cross validation fold. Yaxis, right figure, MeanEuclDistAllFold, Mean Euclidean distance across all outer cross validation folds.



Supplementary Figure S3. Comparison of different methods using organizers scoring functions. For sub-challenge 1, Lasso.TopX performed better than NeuralNets for s1 and s3. For sub-challenge 2, NeuralNets performed better for all scores and for sub-challenge 3 Lasso.TopX performed better for s1 and s3 scores. In all sub-challenges, both methods performed better than Random (Random_selGenes). The binarized expression data that were produced using all expression data, _PB extension, showed an extreme bias in overestimation of performance, across all metrics, methods and sub-challenges. Lasso.TopX_PB performed always better than Lasso.TopX_selGenes, NeuralNets_PB performed always better than NeuralNets_selGenes and Random_PB performed always better than Random_selGenes.



Supplementary Figure S4.
Percentage of gene overlaps between Lasso.TopX and NN across the 10 fold outer cross validation, using either only inSitu genes a), or all genes b). Yaxis: the percent of common genes between Neural Nets and Lasso.TopX, and Xaxis: the number of features selected. Boxplots and the corresponding data points in blue, are shown.



Supplementary Figure S5. Features stability in Outer CV using all genes.
The percentage of common selected genes across the 10 fold outer cross validation is shown. a) In the Lasso.TopX case, 30 genes were selected across all folds for subchallenge 3, red dots. 4 genes or 13.3% percent were selected only in 1 fold, 4 or 13.3% percent were selected by 3 folds, etc. 11 or 36.6% percent were selected in all 10 folds. b) In the NNs case, out of a total of 28 genes selected across all folds for subchallenge 3, red dots, 13 genes or 46.4% were selected across all 10 folds of the outer cross validation. Similar results are shown for subchallenge 1, and subchallenge 2, blue and green dots, respectively.