**Comparison between PSWM and SVM based models for signal peptides detection: SUPPLEMENTARY MATERIALS**

**Datasets statistics**

- SPs length distribution, fig. 1A and 1B

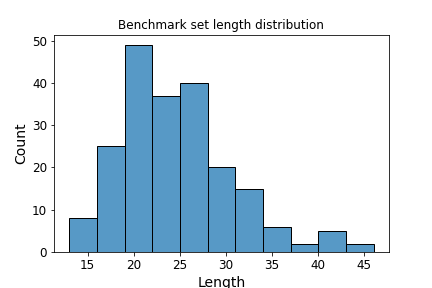
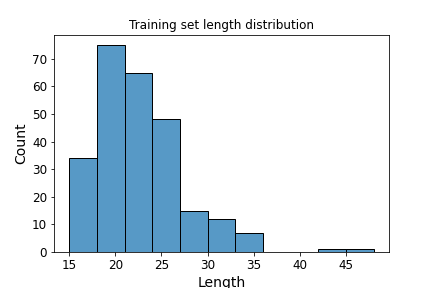


Fig. 1A and 1B | Histograms of the distribution of SP lengths across the SP sequences included in the datasets

- Compositional analysis, fig 2A and 2B

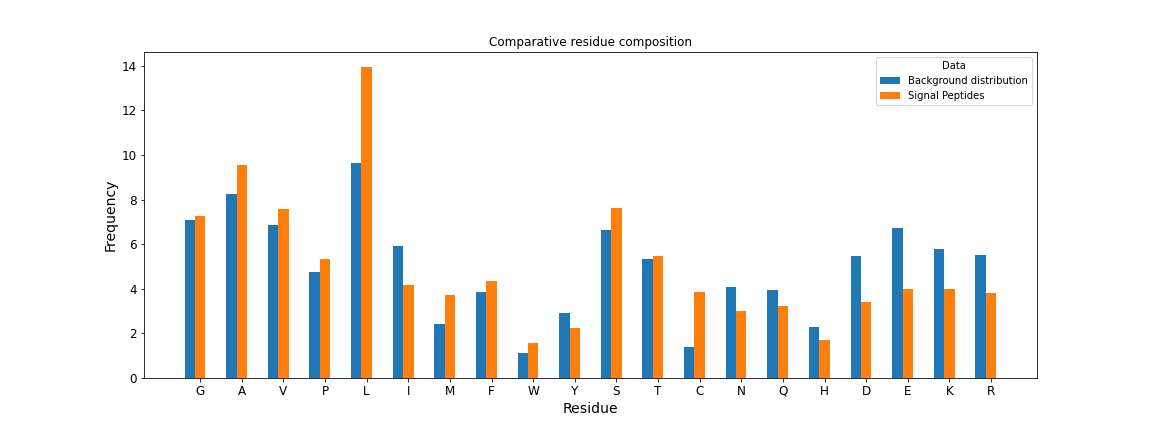


Fig. 2A and 2B | Histograms of the composition of SP sequences included in the datasets. On the y-axis is reported the frequency in percentage (%).

- Taxonomic analysis

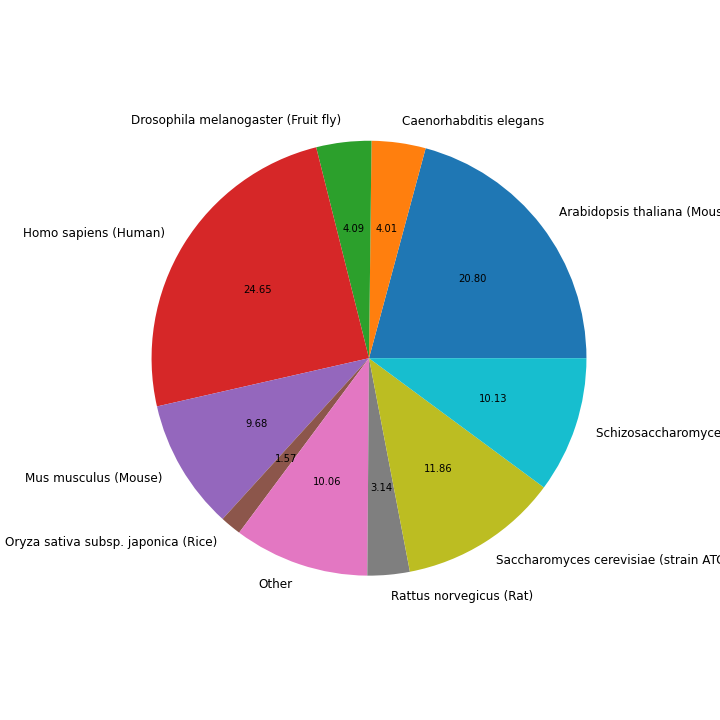
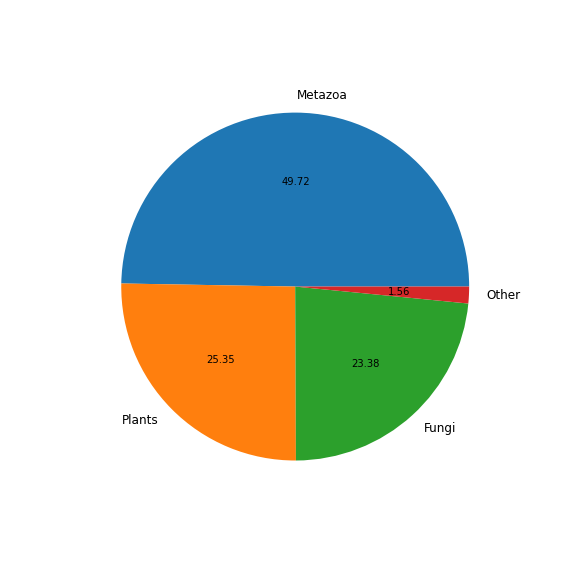
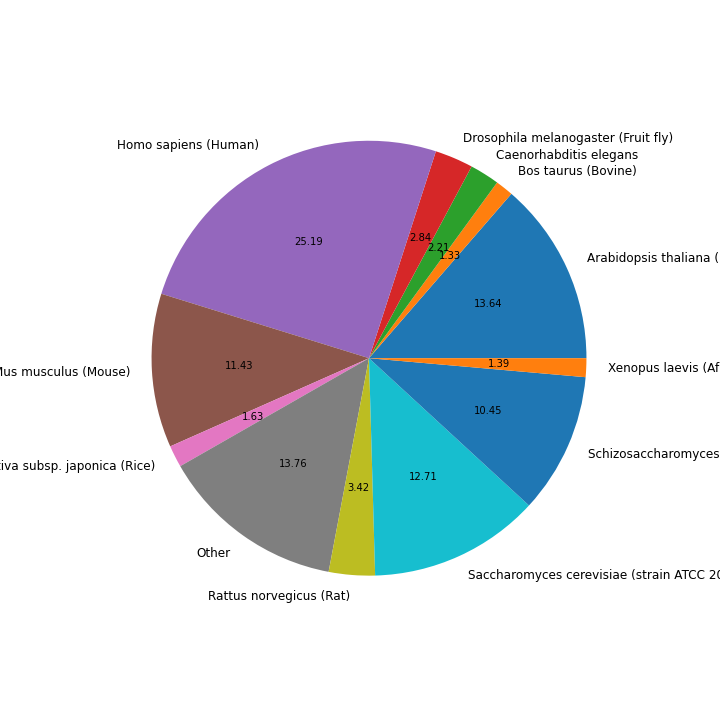
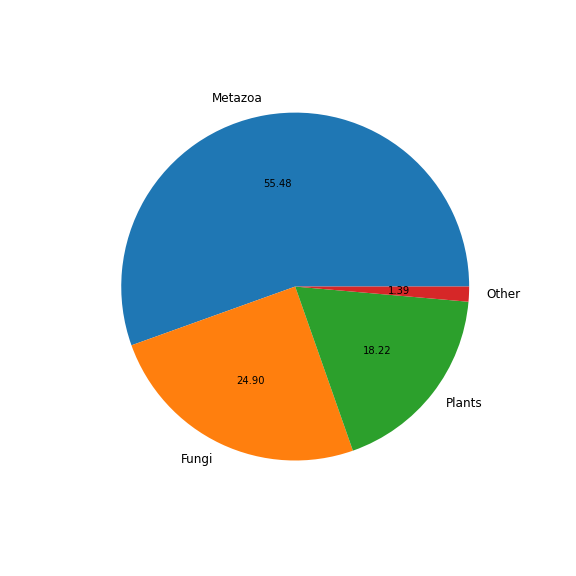
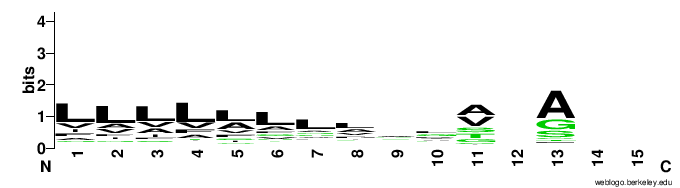


Fig. 3A and 3B | Piecharts reporting the SP sequences distribution in percentage across kingdoms (up) and taxa (down) for training set (left) and for benchmark set (right).

- Sequence Logo



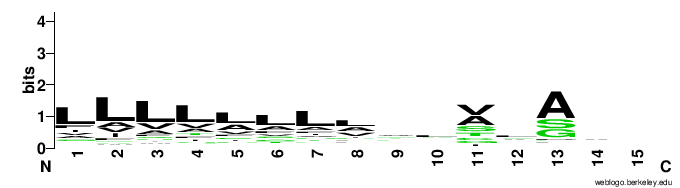


Fig. 4A and 4B | Sequence logo for training set (up) and for benchmark set (down).

**Cross-validation results, von Heijne model**

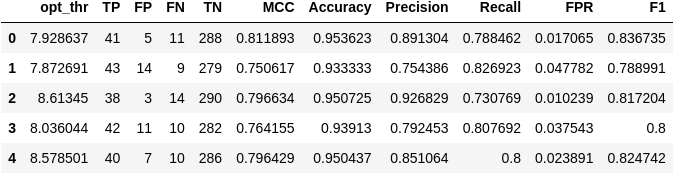
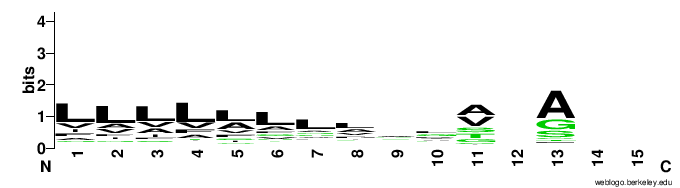
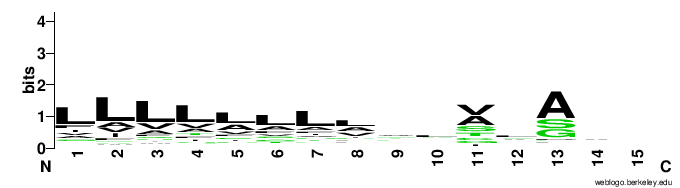


Table A | Report of the 5-fold cross-validation runs for the von-Heijne model. The final threshold was taken as the mean vaue of the thresholds.

**False negatives analysis**

- von-Heijne



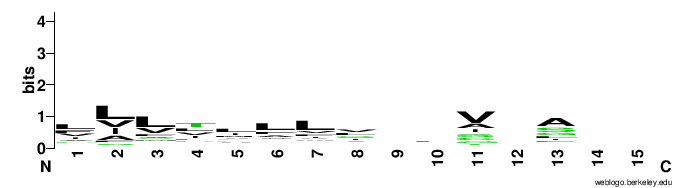


Fig. 5A | Sequence logo for the false negative results of the von-Heijne model (down). For visual comparison are reported again the Logo of training set (up, left) and for benchmark set (up, right). To be noticed the difference in composition between training set and false results.

- SVM

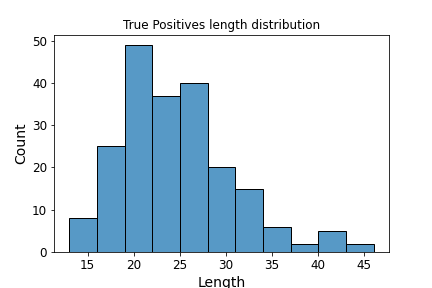
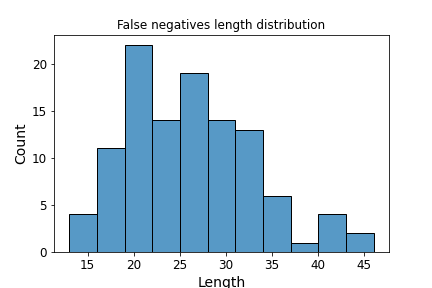


Fig. 5B | Histogram of the lengths distribution across the FN sequences. For visual comparison, is reported again the length distribution of the benchmark set positives.

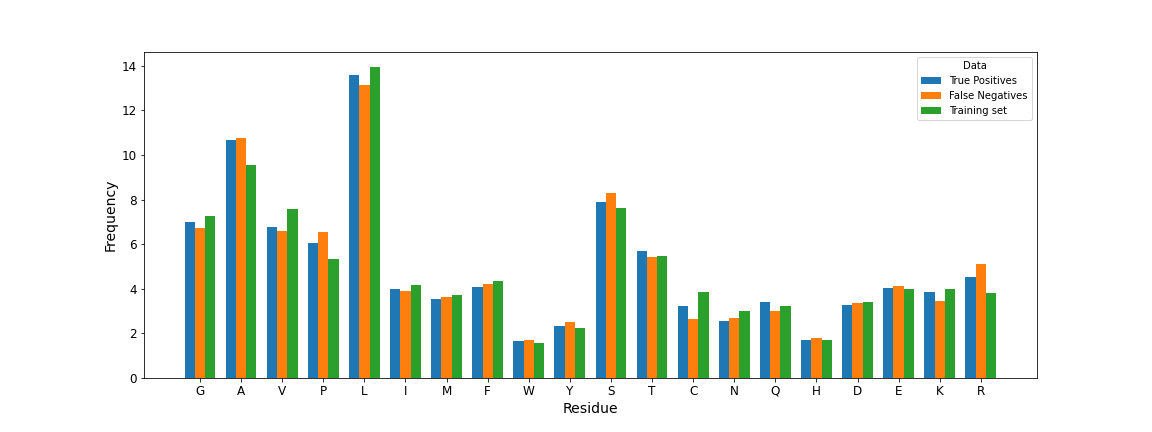


Fig. 5C | Histograms of the composition of false negatives sequences from the benchmark set and SP sequences from benchmark (i.e. True Positives) and training datasets. On the y-axis is reported the frequency in percentage (%).