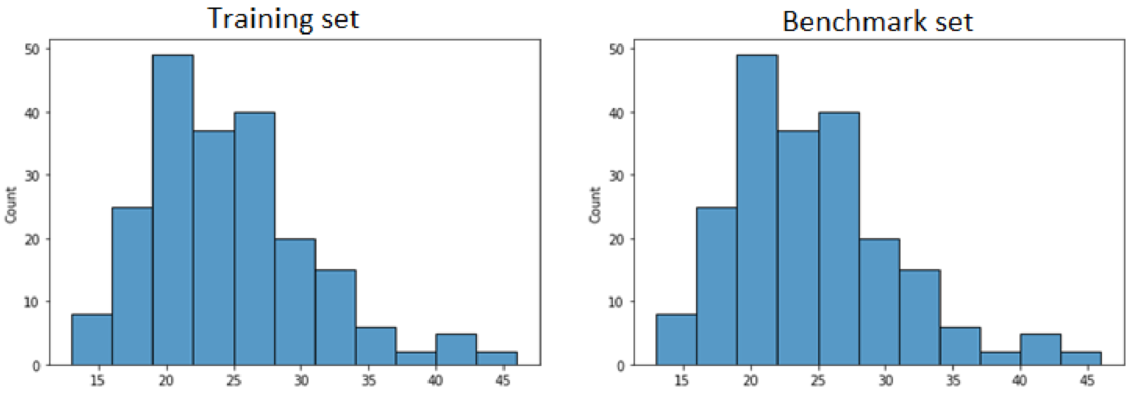
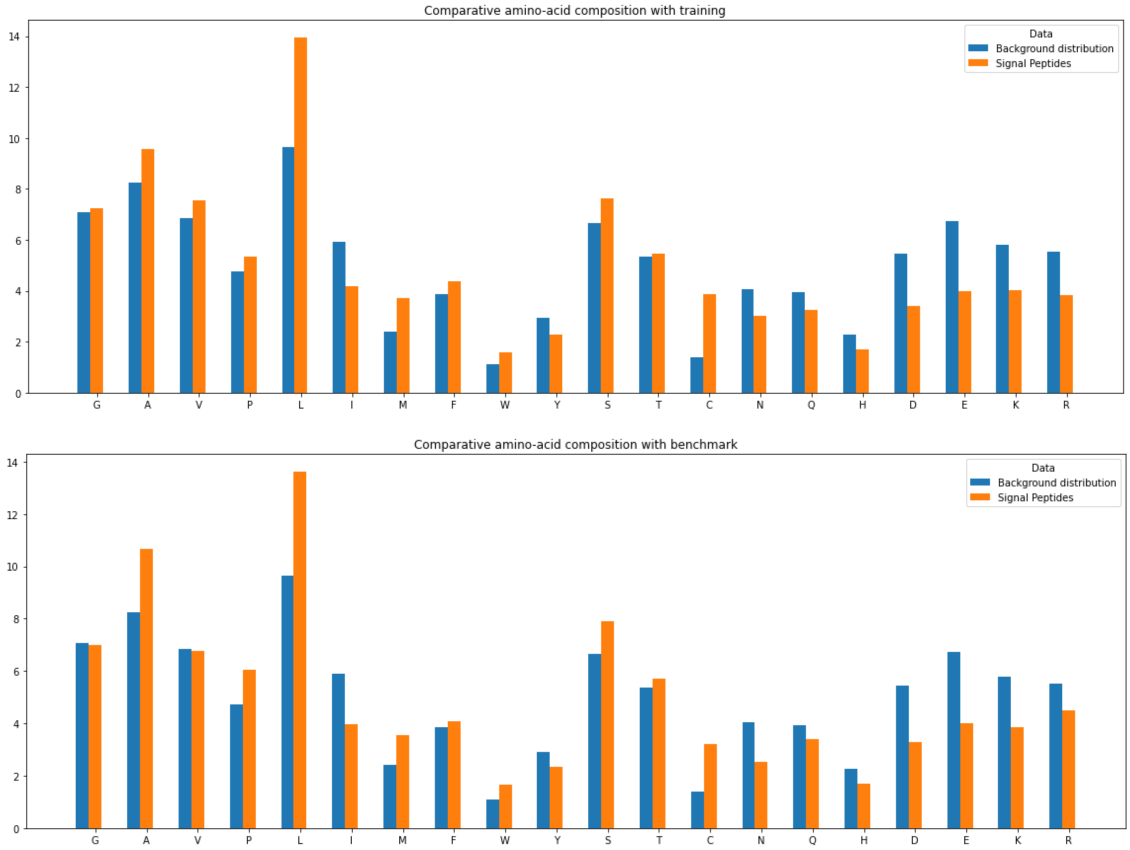
Supplementary materials for “**Build of a model for the detection of signal peptides in proteins**”

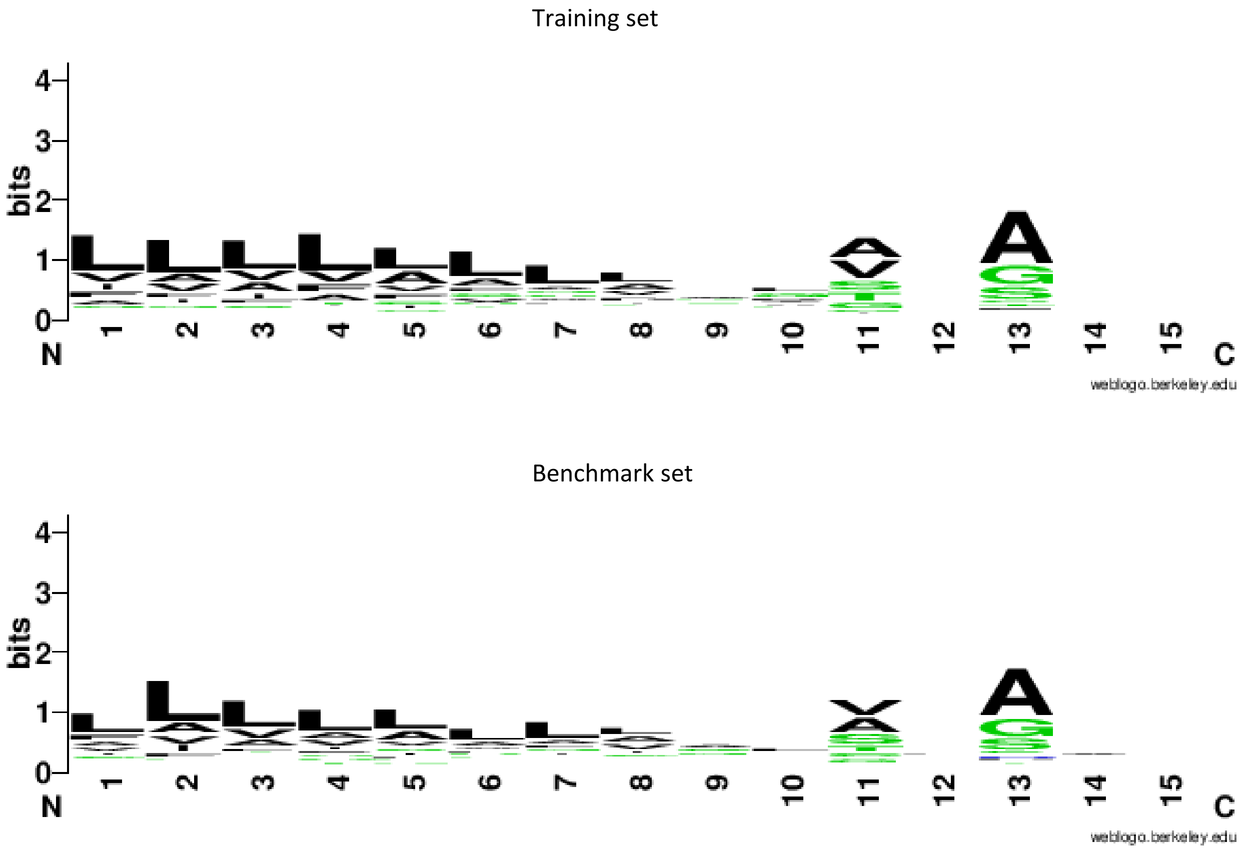
Paolo Mastrogiovanni



Supplementary Figure 1 – Histograms showing the distribution of the signal peptides lengths in the training and benchmark dataset.



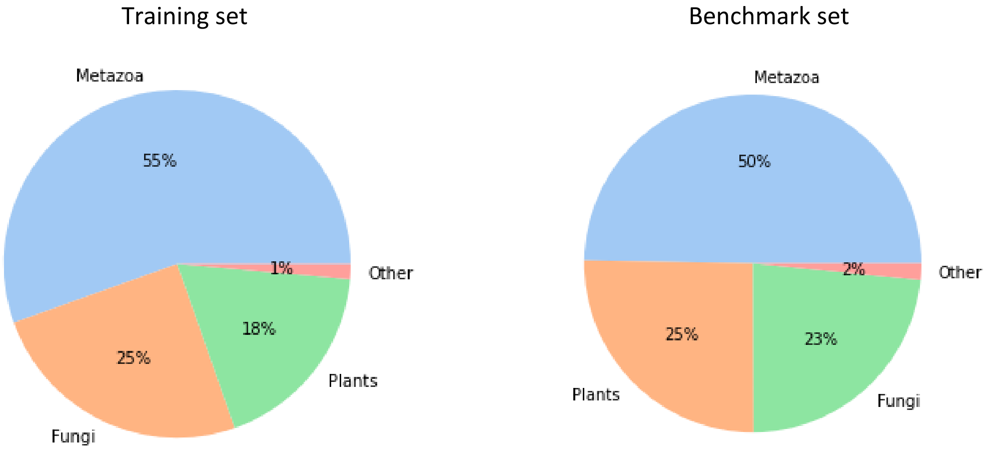
Supplementary Figure 2 – Comparation between the aminoacidic composition of sequences from the training and benchmark dataset with the background distribution (Swissprot). Signal peptides are characterized by a larger amount of hydrophobic residues.



Training set

Benchmark set

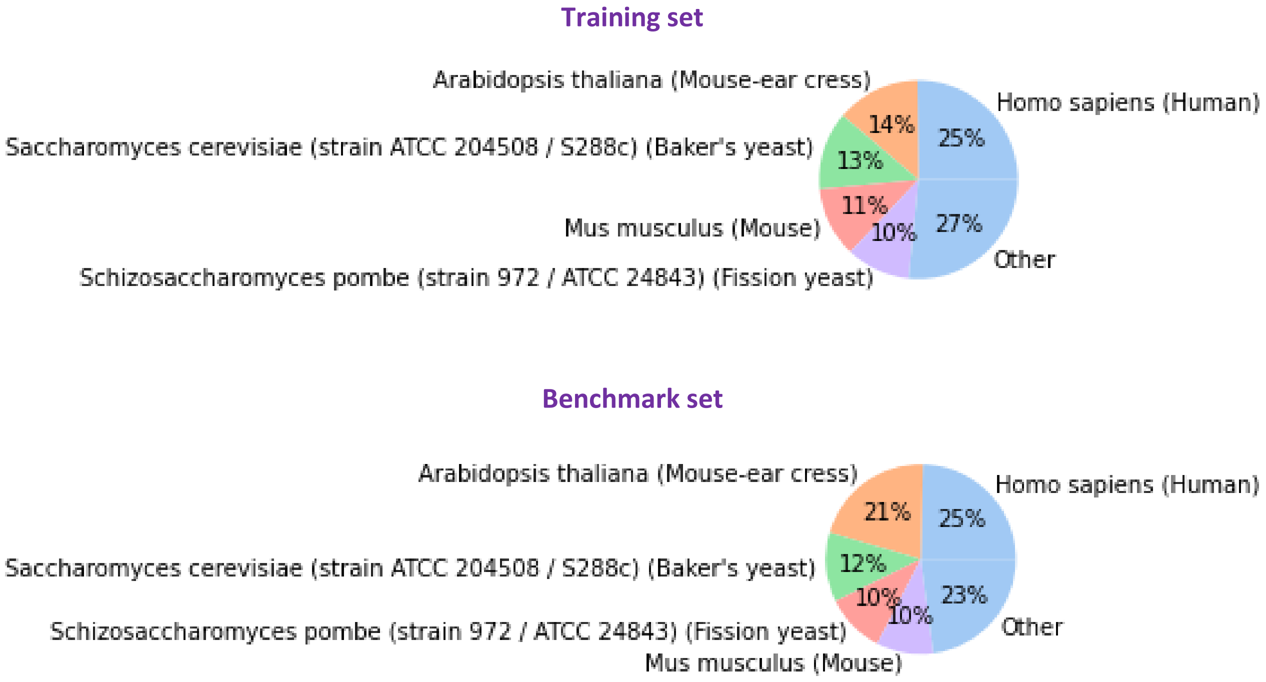
Supplementary Figure 3 – Sequence Logo computed for the sequences of the signal peptides in the training set and in the benchmark set. It is a graphical representation of the sequence conservation of nucleotides.



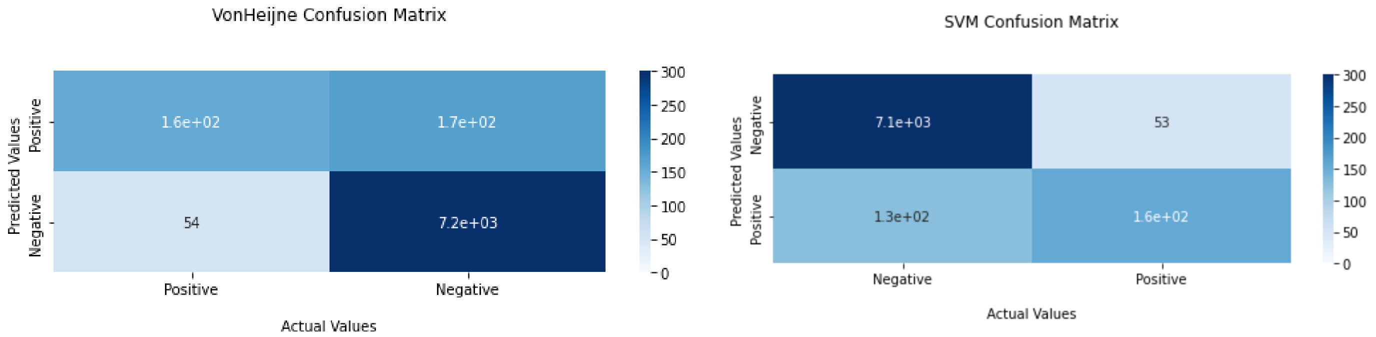
Training set

Benchmark set

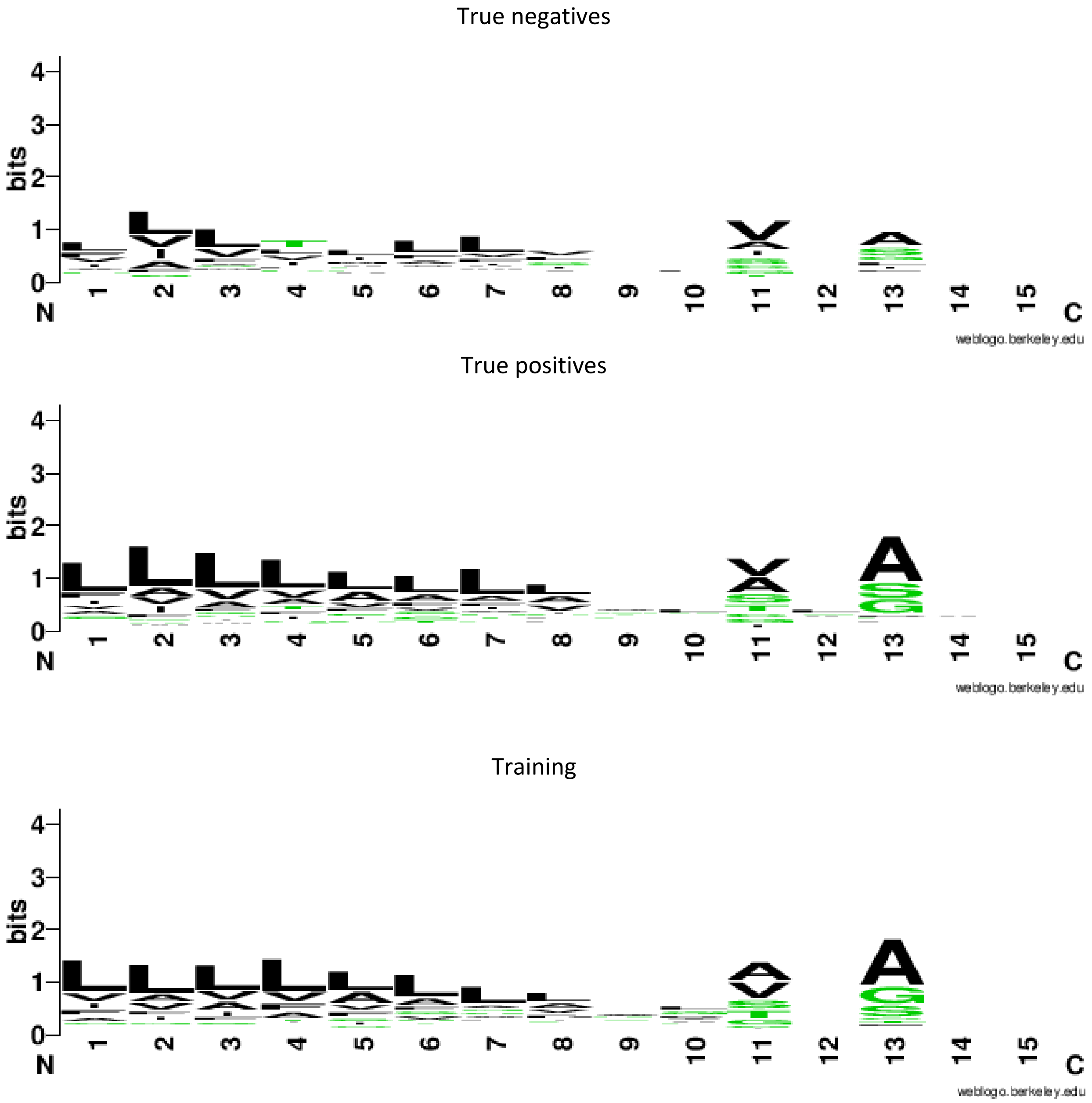
Supplementary Figure 4.1 – Distribution of the signal peptides sequences of training and benchmark set among Kingdoms. Most of the proteins can be identified in the Metazoa kingdom.



Supplementary Figure 4.2 – Distribution of the signal peptides sequences of training and benchmark set among Taxa. Sequences are well distributed among taxas.



Supplementary Figure 5 – Confusion matrices produced with Seaborn for both the Von Heijne method and the SVM testing.

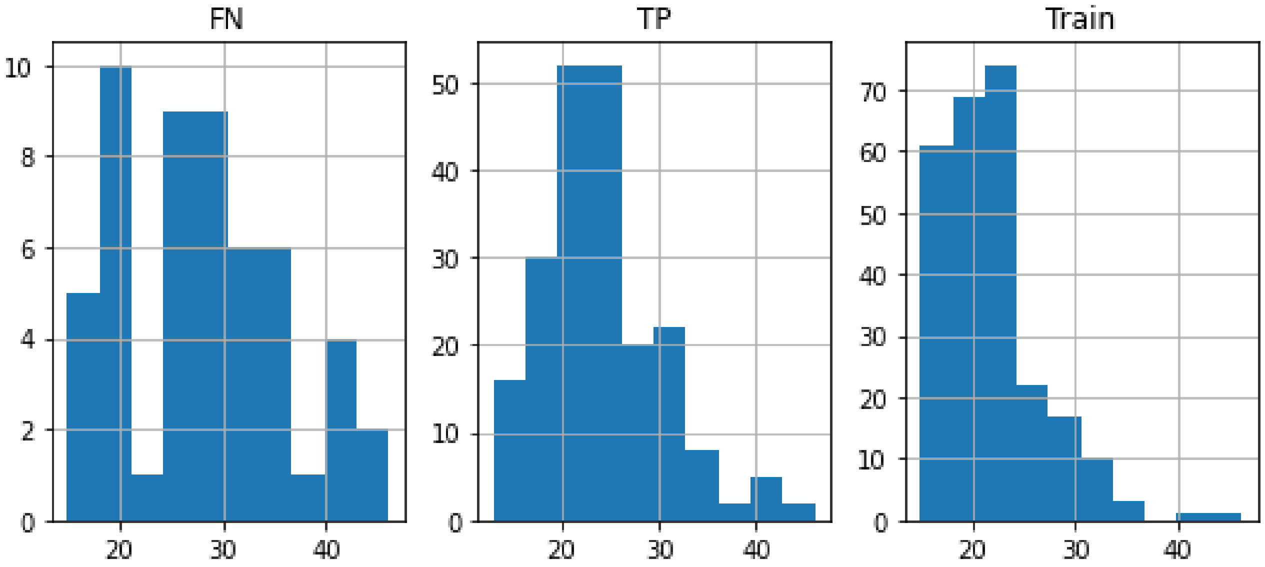


True negatives

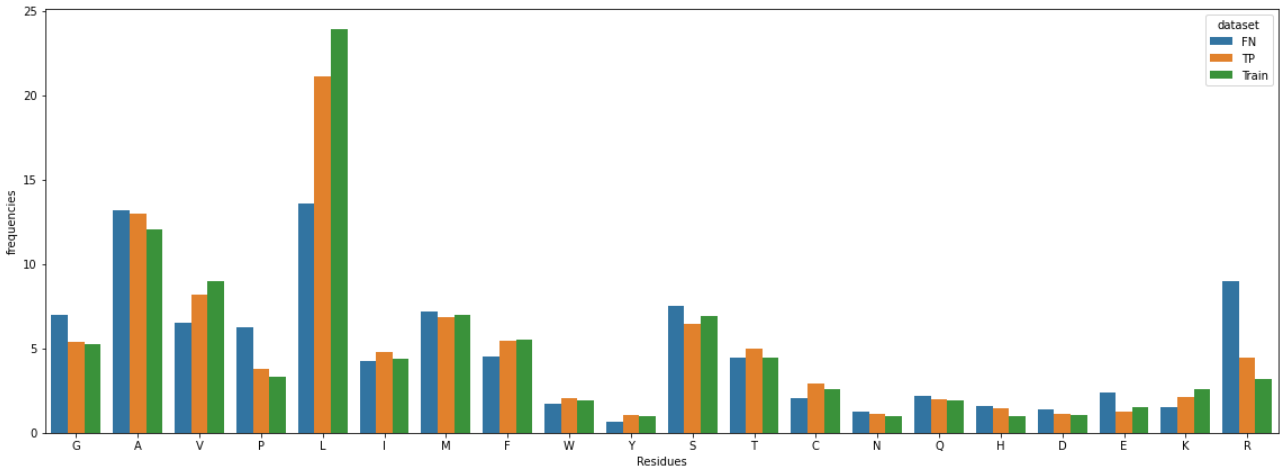
True positives

Training

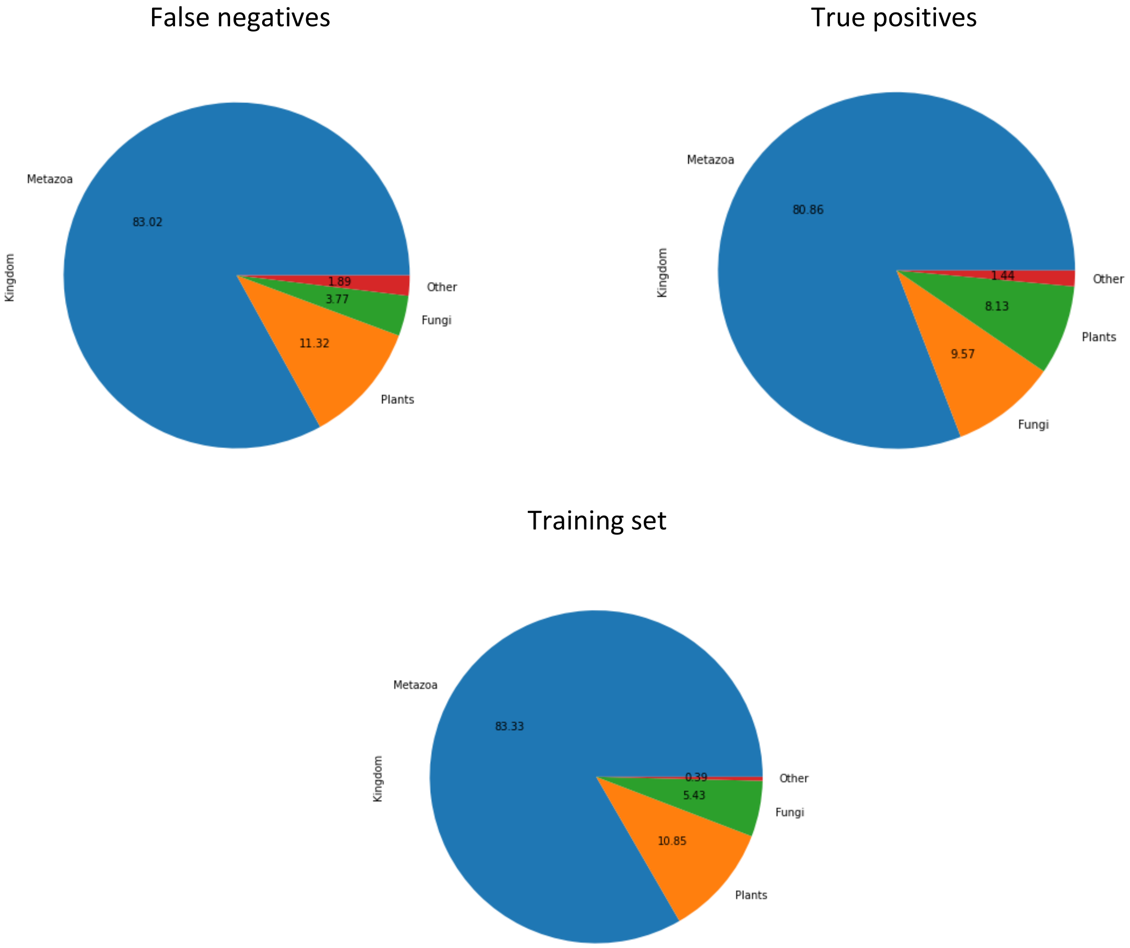
Supplementary Figure 6 – Wrong prediction analysis for the Von Heijne model. The image shows a comparison between the sequence logos of the signal peptides cleavage site for the false negatives, true positives, and training set. The correctly predicted entries have a very similar composition to the training set, while the false negatives have many differences, especially in the AXA motif, which is way less represented.



Supplementary Figure 7 – Wrong prediction analysis for the SVM. Signal peptides length distribution, comparison between the false negatives, the true positives (from the benchmark set) and the training set positives.



Supplementary Figure 8 – Wrong prediction analysis for the SVM. Comparison of the aminoacidic composition of the signal peptide sequences between false negatives, true positives, and training set positives. Main differences between the FNs set and the other proteins lies in a less represented leucine, and a more represented arginine.



Supplementary Figure 9 – Distribution of the sequences among kingdoms for the false negatives of the svm, compared with true positives and training set positives.