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

Datasets statistics

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

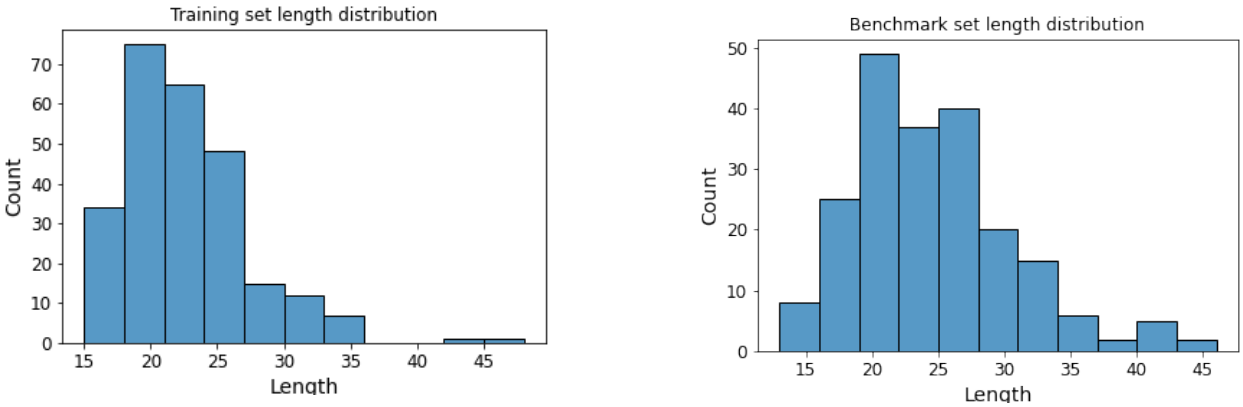


Fig. 1A and 1B | Histograms of the distribution of SP lengths across the SP sequences included in the datasets

- Compositional analysis, fig 2A and 2B

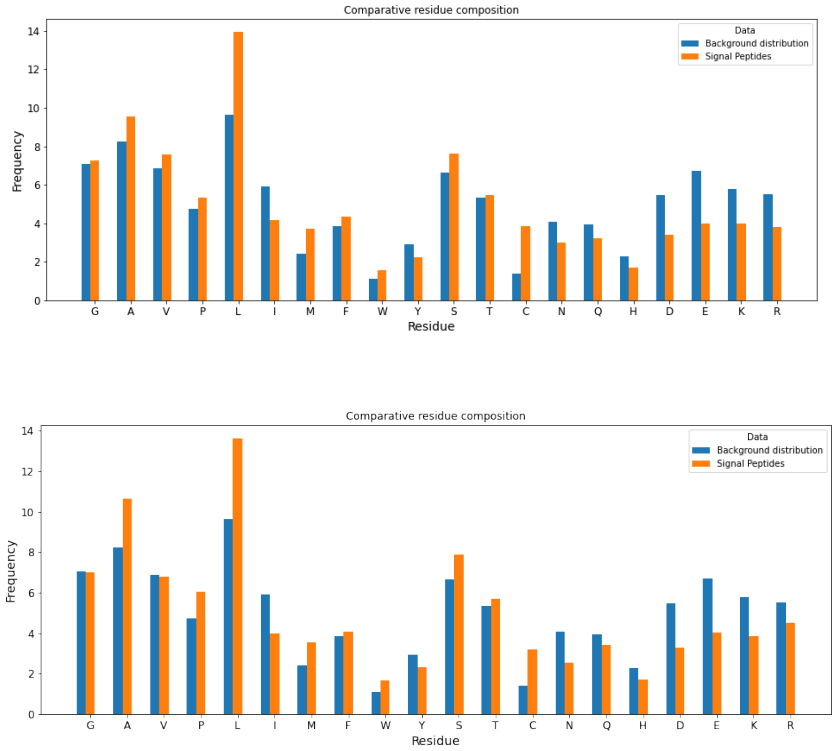


Fig. 2A and 2B | Histograms of the composition of SP sequences included in the datasets. On the y-axis is reported the frequency in percentage (%).

- Taxonomic analysis

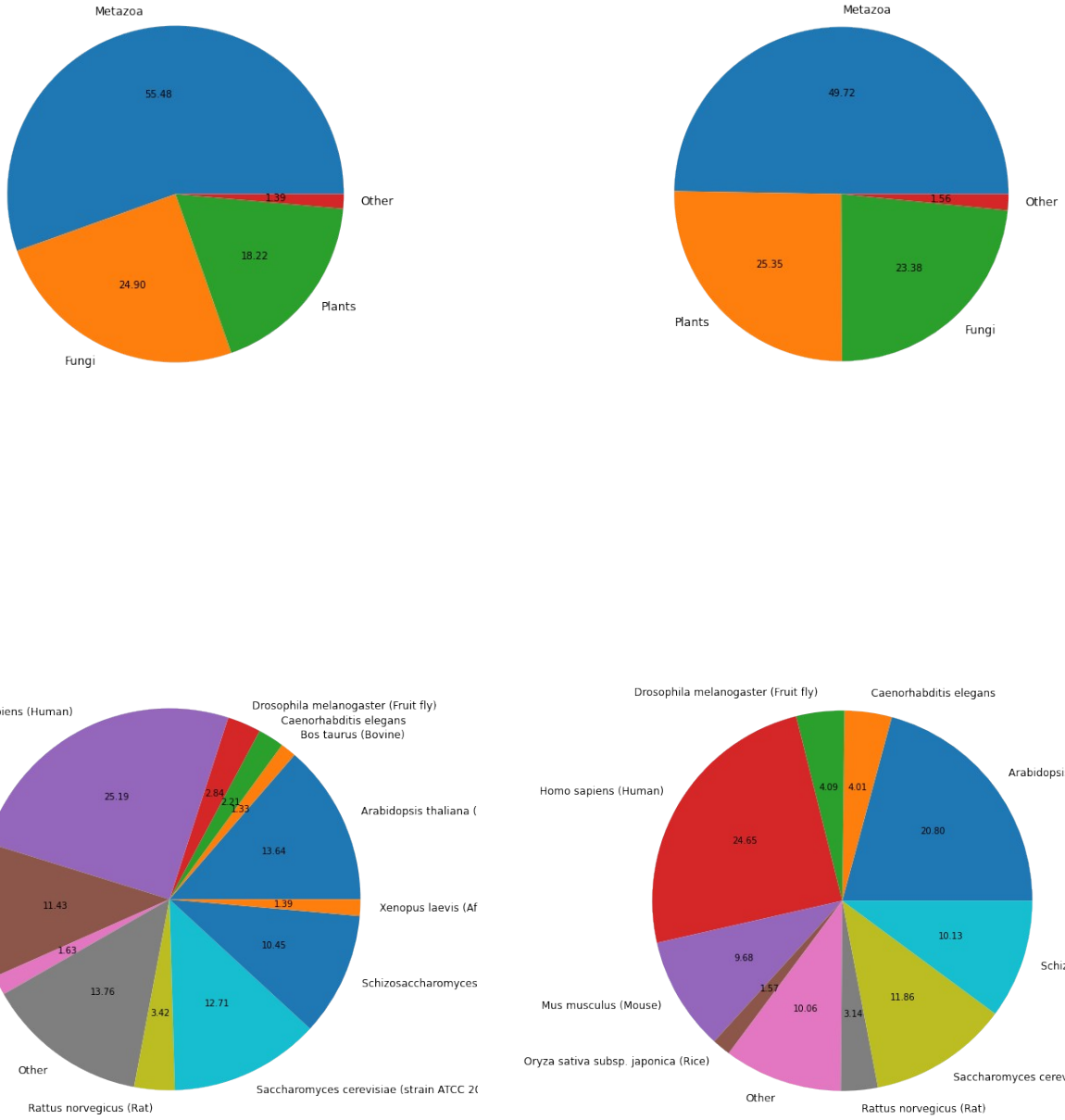


Fig. 3A and 3B | Piecharts reporting the SP sequences distribution in percentage across kingdoms (up) and taxa (down) for training set (left) and for benchmark set (right).

- Sequence Logo

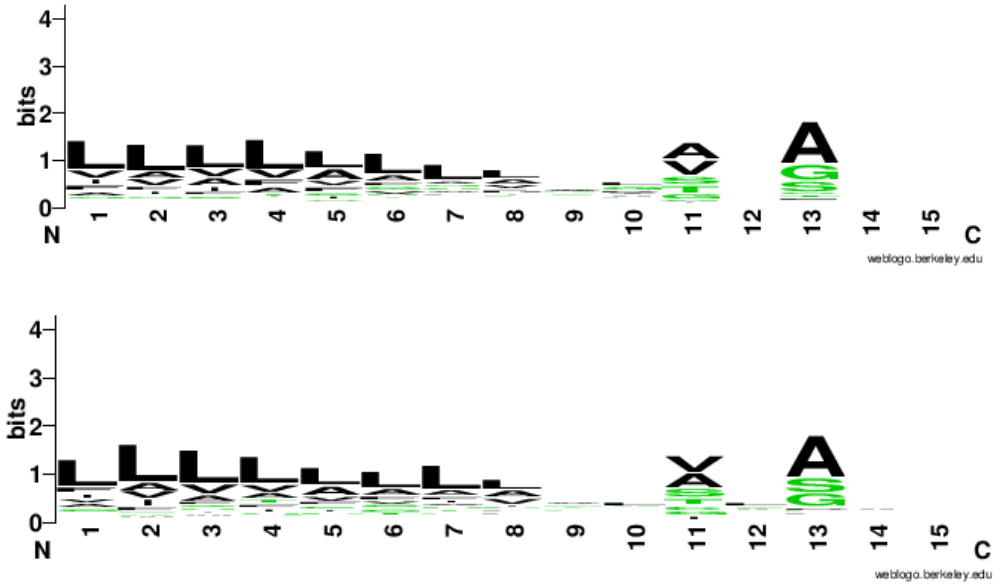


Fig. 4A and 4B | Sequence logo for training set (up) and for benchmark set (down).

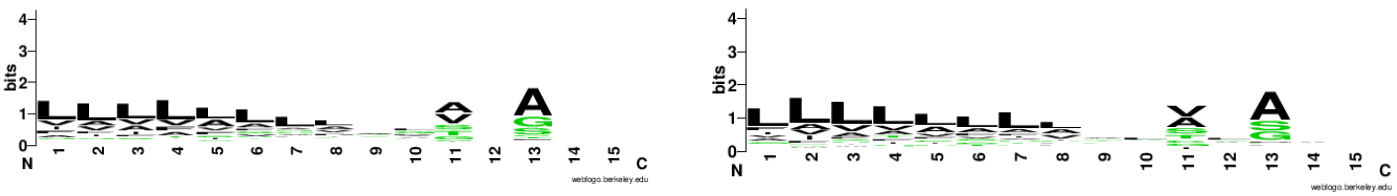
Cross-validation results, von Heijne model

	opt_thr	TP	FP	FN	TN	MCC	Accuracy	Precision	Recall	FPR	F1
0	7.928637	41	5	11	288	0.811893	0.953623	0.891304	0.788462	0.017065	0.836735
1	7.872691	43	14	9	279	0.750617	0.933333	0.754386	0.826923	0.047782	0.788991
2	8.61345	38	3	14	290	0.796634	0.950725	0.926829	0.730769	0.010239	0.817204
3	8.036044	42	11	10	282	0.764155	0.93913	0.792453	0.807692	0.037543	0.8
4	8.578501	40	7	10	286	0.796429	0.950437	0.851064	0.8	0.023891	0.824742

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



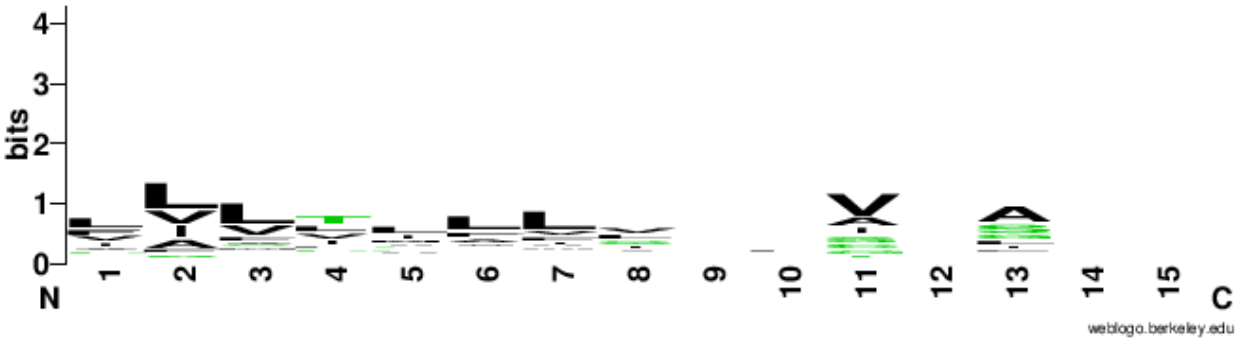


Fig. 5A | Sequence logo for the false negative results of the von-Heijne model (down). For visual comparison are reported again the Logo of training set (up, left) and for benchmark set (up, right). To be noticed the difference in composition between training set and false results.

- SVM

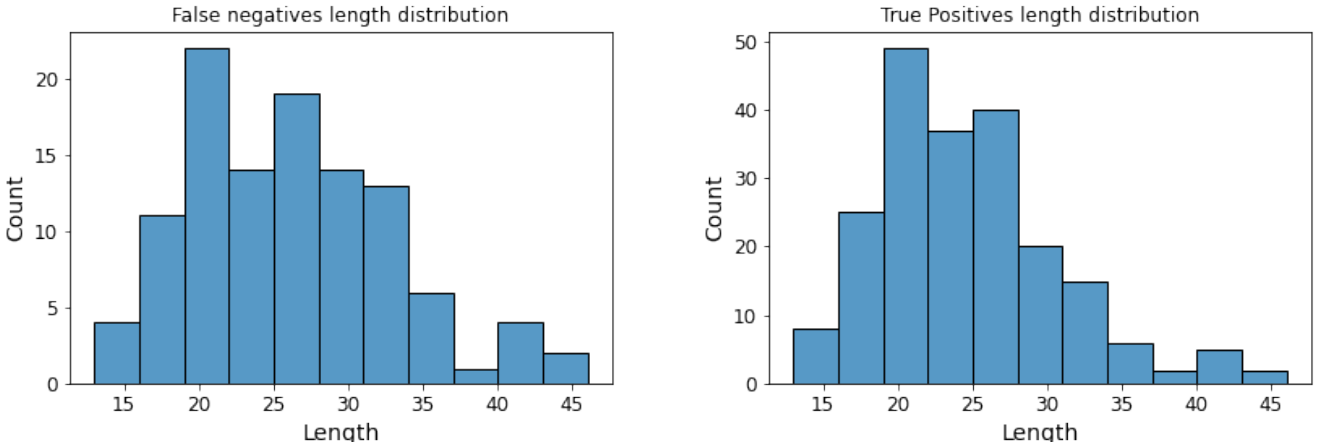


Fig. 5B | Histogram of the lengths distribution across the FN sequences. For visual comparison, is reported again the length distribution of the benchmark set positives.

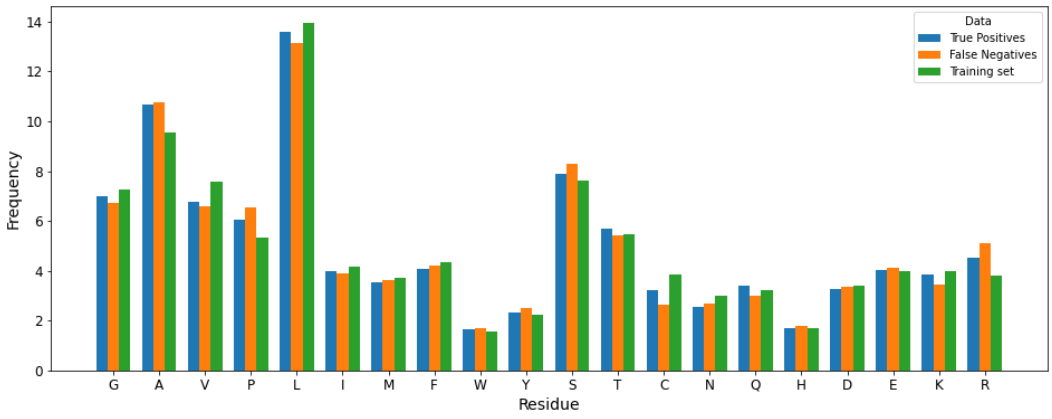


Fig. 5C | Histograms of the composition of false negatives sequences from the benchmark set and SP sequences from benchmark (i.e. True Positives) and training datasets. On the y-axis is reported the frequency in percentage (%).