Supplementary materials

Models for signal peptide prediction: comparing Von Heijne algorithm and Support Vector Machines

Giacomo Orsini [1,*]

[1] Department of Pharmacology and Biotechnology, Alma Mater Studiorum - Università di Bologna, Bologna, Italy

* To whom correspondence should be addressed.

1 Supplementary Images

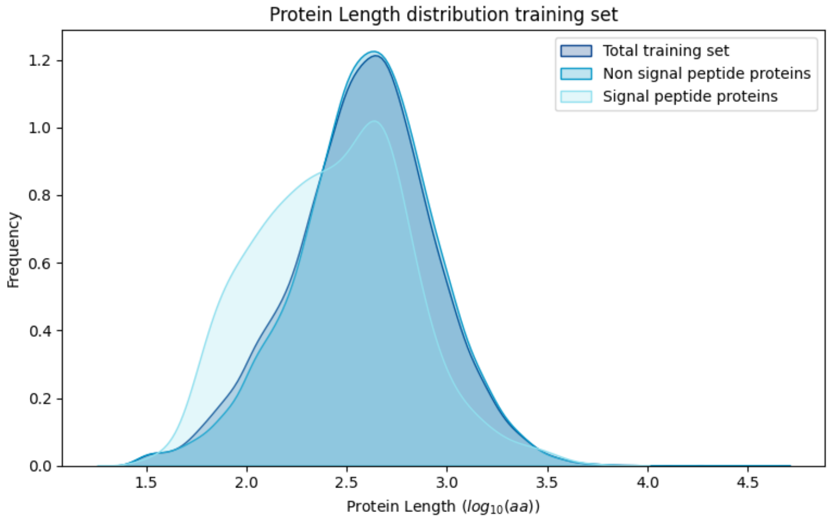

Fig. 1. Normalized length distribution for the proteins of the Training set. A difference in the distribution shape can be appreciated between positive and negative proteins.

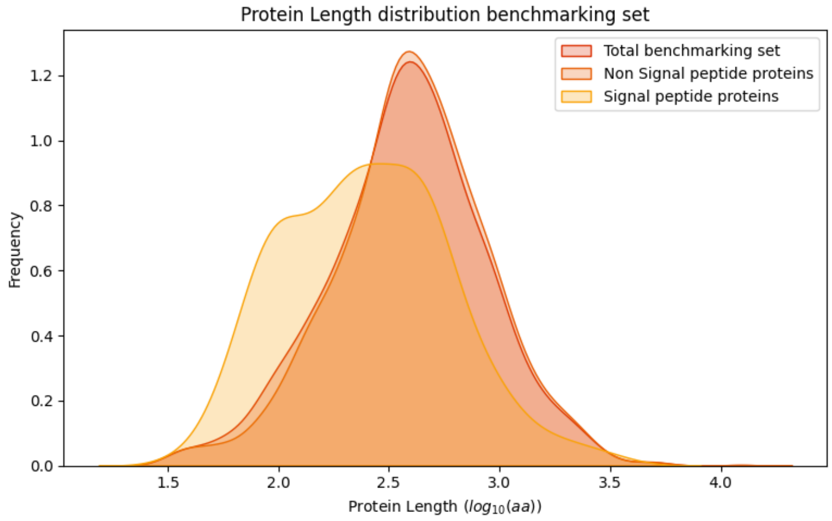

Fig. 2. Normalized length distribution for the proteins of the Benchmarking set. A difference in the distribution shape can be appreciated between positive and negative proteins.

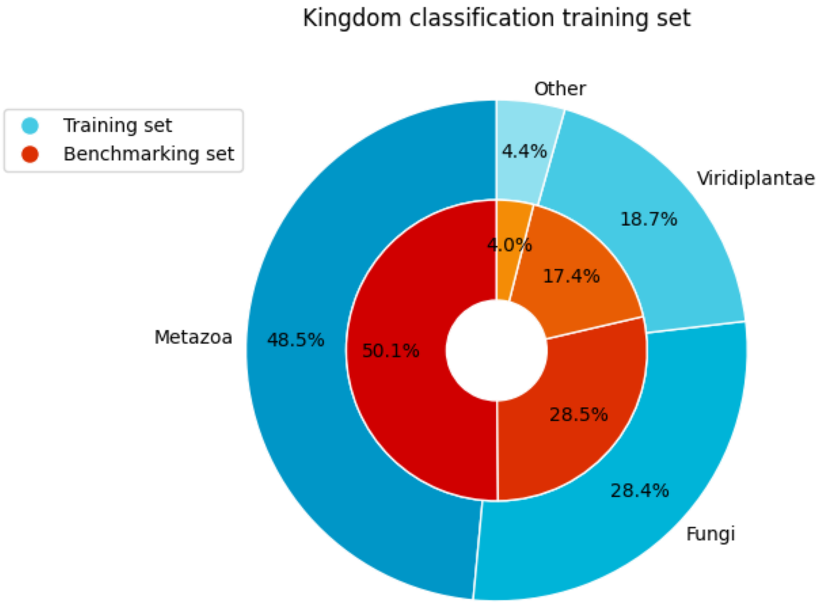

Fig. 3. Classification at kingdom level for proteins of Training and Benchmarking set. Unknown taxa were defined as "Others"

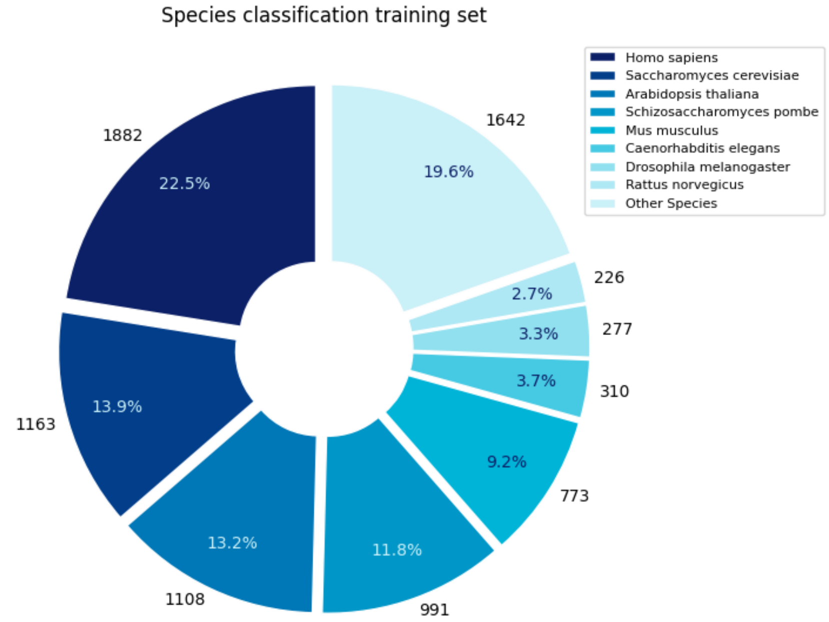

Fig. 4. Classification at species level of the proteins of the training set. The most frequent 8 species are shown, the others are clustered together.

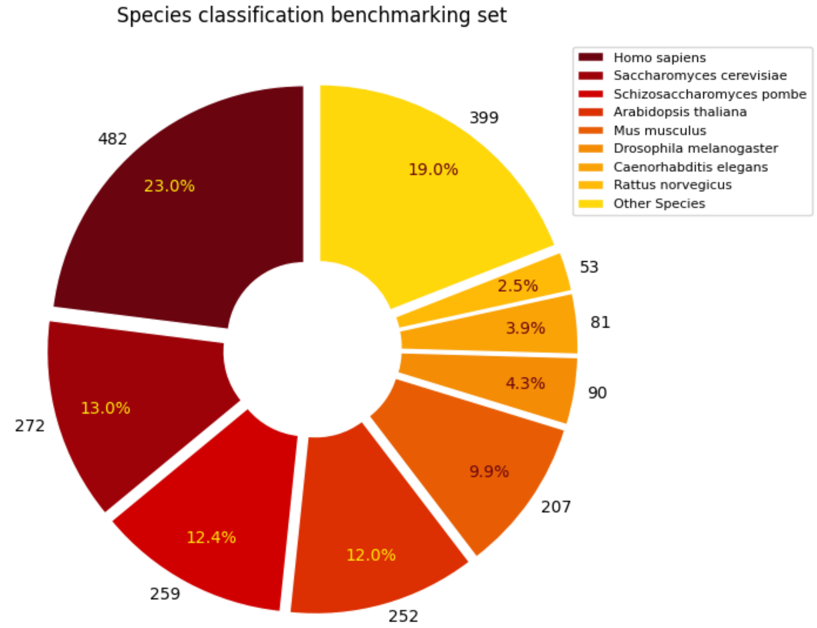

Fig. 5. Classification at species level of the proteins of the Benchmarking set. The most frequent 8 species are shown, the others are clustered together.

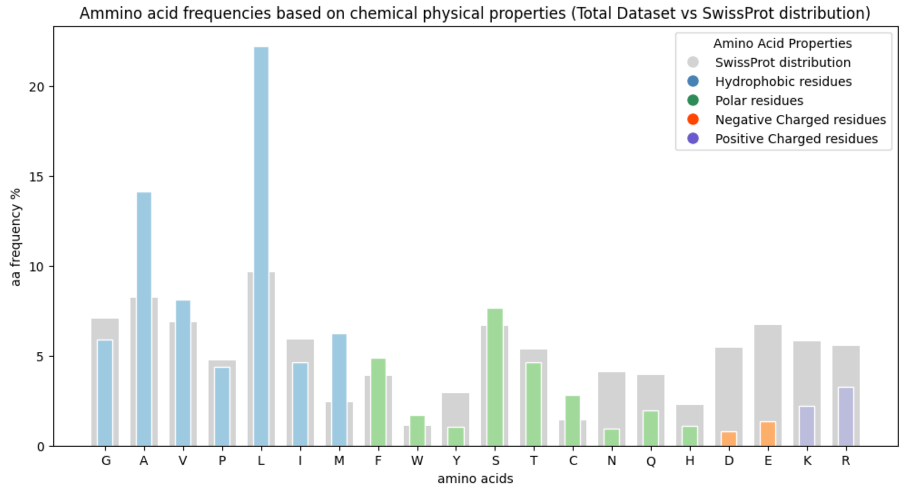

Fig. 6. Amino acidic frequencies with highlightened amino acidic properties. In background the SwissProt distribution; the benchmarking and training distributions have been united together.

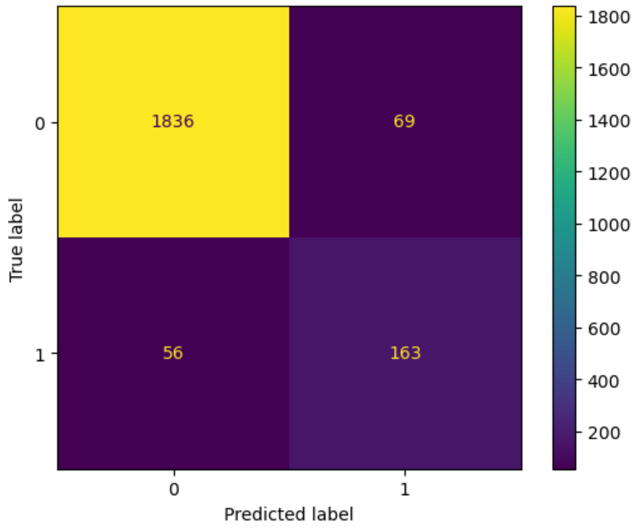

Fig. 7. Confusion matrix of the results of the final Von Heijne model

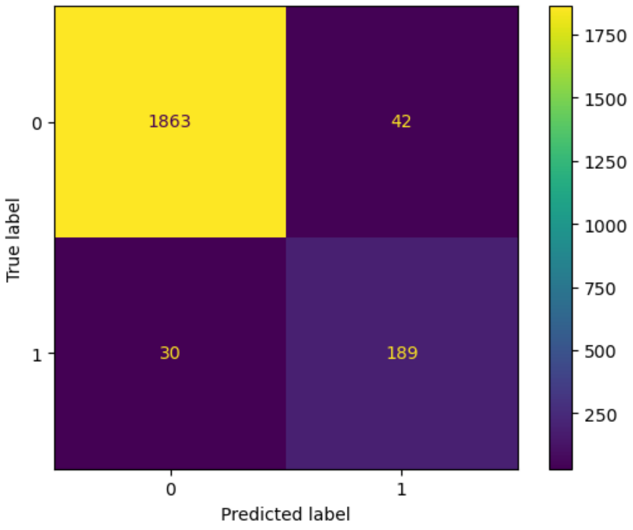

Fig. 8. Confusion matrix of the results of the final SVM model

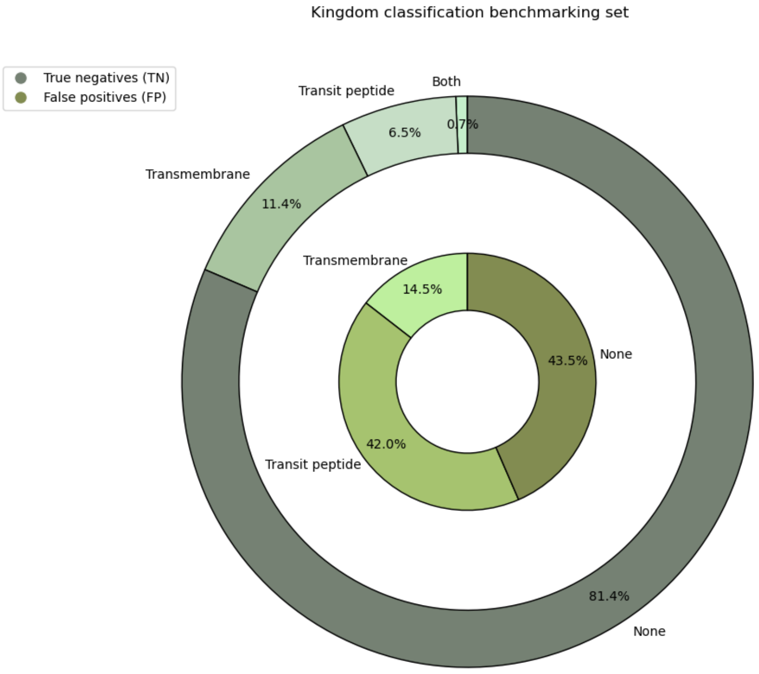

Fig. 9. Percentage of Transit peptides, Transmembrane proteins, proteins categorized as both and as none of the two of the False positives subset (benchmarking VH) and True negatives subset (benchmarking VH). The relative frequency of TP and TM is greater in the False positives subsets.

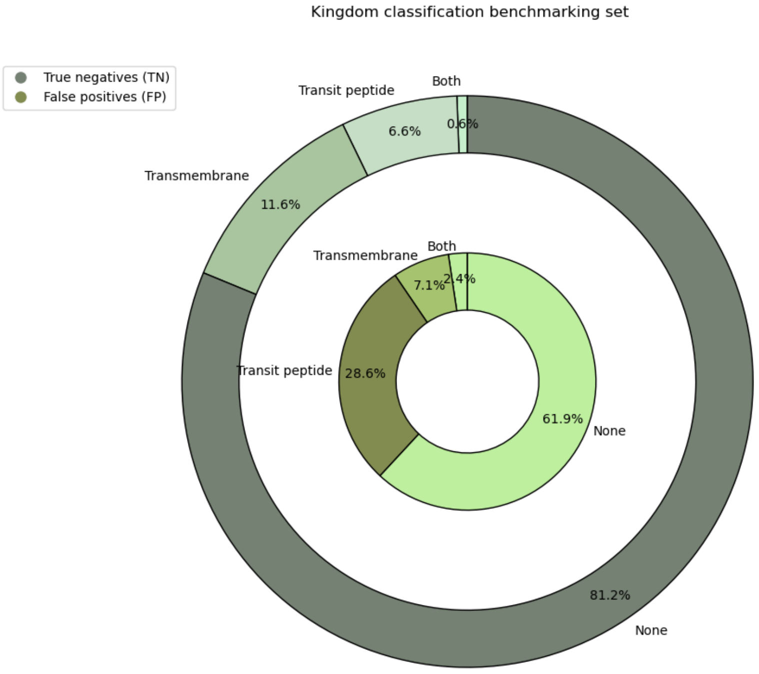

Fig. 10. Percentage of Transit peptides, Transmembrane proteins, proteins categorized as both and as none of the two of the False positives subset (benchmarking SVM) and True negatives subset (benchmarking SVM). The relative frequency of TP and TM is greater in the False positives subsets.

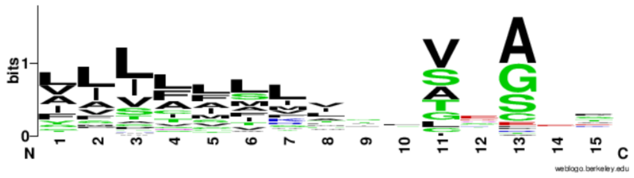

Fig. 11. Sequence logo of the cleavage sites of the False negatives entries in the benchmarking set, considering the results of the Von Heijne final model.

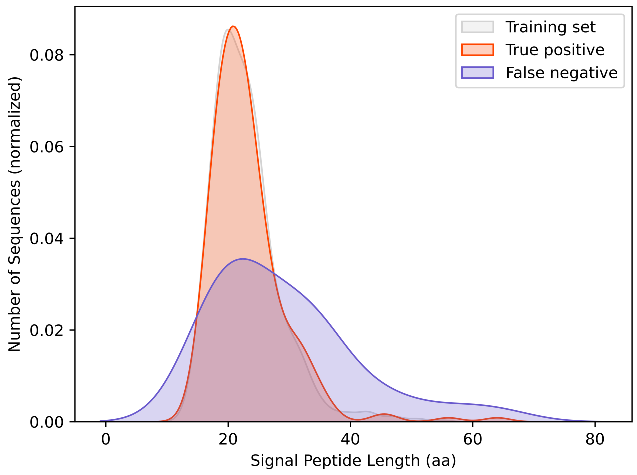

Fig. 12. Signal peptide length distributions for the False negatives subset and True positive subset of the benchmarking set, and the positives proteins of the training set, considering the results of the final SVM model.

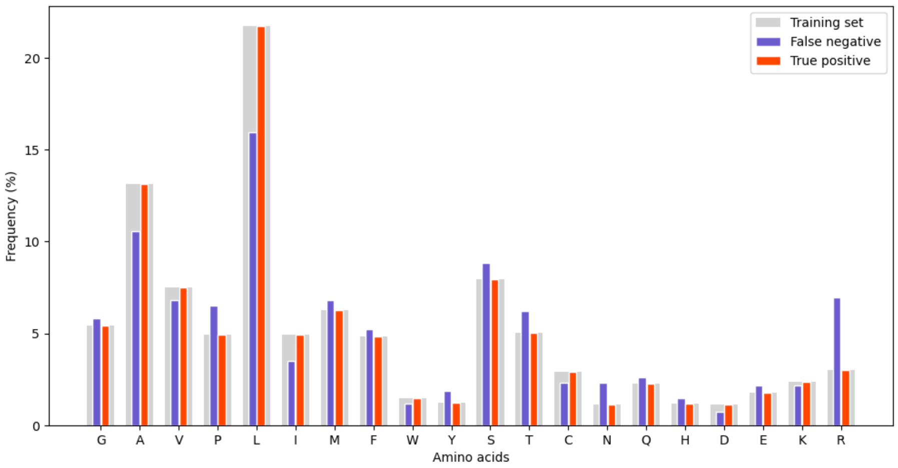

Fig. 13. Amino acidic distribution of the first 23 residues for the False negatives subset and True positive subset of the benchmarking set, and the positives proteins of the training set, considering the results of the final SVM model.

2 Tables

Table 1.

Metric	Value
total count	874
mean length	22.862
minimal length	13
maximal length	55

Statistical analysis on SP length of training set positive proteins.

Table 2.

Matric	Value
total count	219
mean length	23.151
minimal length	15
maximal length	64

Statistical analysis on SP length of benchmarking set positive proteins.

Table 3.

Metric	Value
total count	7618
mean length	546.043
minimal length	30
maximal length	34350

Statistical analysis on protein length of training set proteins.

Table 4.

Metric	Value
total count	874
mean length	398.550
minimal length	35
maximal length	5263

Statistical analysis on protein length of benchmarking set proteins

Table 5.

Run	Training set	Validation set	Testing set	Threshold	MCC	F1	ACC	Precision	Recall	AUC	TN	FP	FN	TP
1	CV2 +, CV3 +, CV4 +	CV1	CV0	6.085	0.627	0.742	0.936	0.67	0.747	0.964	1460	64	44	130
2	CV3 +, CV4 +, CV0 +	CV2	CV1	5.816	0.649	0.694	0.929	0.629	0.754	0.955	1446	78	43	132
3	CV4 +, CV0 +, CV1 +	CV3	CV2	6.398	0.658	0.707	0.934	0.665	0.726	0.955	1460	64	48	127
4	CV0 +, CV1 +, CV2 +	CV4	CV3	6.554	0.723	0.757	0.951	0.783	0.72	0.954	1489	35	49	126
5	CV1 +, CV2 +, CV3 +	CV0	CV4	6.376	0.695	0.731	0.944	0.733	0.72	0.967	1476	46	49	126

Extended results of the cross validation procedure of the Von Heijne method. Performance of each model.

Table 6.

Model (features)	Average MCC	Standard error	K	y	C
C	0.76	±0.017	20	2	2
HP, C	0.809	±0.01	22	2	4
CH, C	0.782	±0.011	20	1	8
AH, C	0.776	±0.011	20	1	8
HP, CH, C	0.81	±0.009	23	1	4
HP, AH, C	0.803	±0.009	23	2	4
CH, AH, C	0.795	±0.007	20	2	8
HP, CH, AH, C	0.809	±0.011	23	scale	8

Extended results of the cross validation procedure of the SVM method. Performance of each model has been obtained by averaging the MCC of the 5 CV sets combination and the hyperparameters choosen have been the most frequent ones.

Table 7.

Run	Training set	Testing set	Threshold	MCC	F1	ACC	Precision	Recall	AUC	TN	FP	FN	TP	
Final	Positive Training set	Benchmarking set	6.246	0.69	0.739	0.742	0.941	0.703	0.744	0.941	1863	69	56	163

Extended performance evaluation of the final Von Heijne model

Table 8.

Run	Training set	Features	K	γ	C	Testing set	MCC	ACC	Precision	Recall	TN	FP	FN	TP
Final	Training set	C, HP, CH	23	1	4	Benchmarking set	0.821	0.966	0.818	0.863	1862	42	30	189

Extended performance evaluation of the final SVM model