Supplementary material

Signal peptide prediction in eukaryotes: a comaparison of SVM performance with a PSWM based method

Mario Esposito[1,*]

[1]Department of Pharmacy and Biotechnology, University 1of Bologna, 40126 Bologna, Italy

*To whom correspondence should be addressed.

Tables

Supplementary Table S1. The whole dataset table is available as file (dataset.tsv). UniProt ID, class, set, cv_subset and sequence are reported.

Supplementary Table S2. Hyperparameters lists adopted in the Grid search CV (SVM). (330 combinations)

Hyperparameter	Values
C	[1,2,4,8,16]
γ	[1,2,4,8,16,'scale'] ('scale' = 1 / [num. of features * global variance])
K	[18,19,20,21,22,23,24,25,26,27,28]

Supplementary Table S3. Avg. feature importance ranking computed among 5 partially overlapping ranking obtained by PFI in CV.

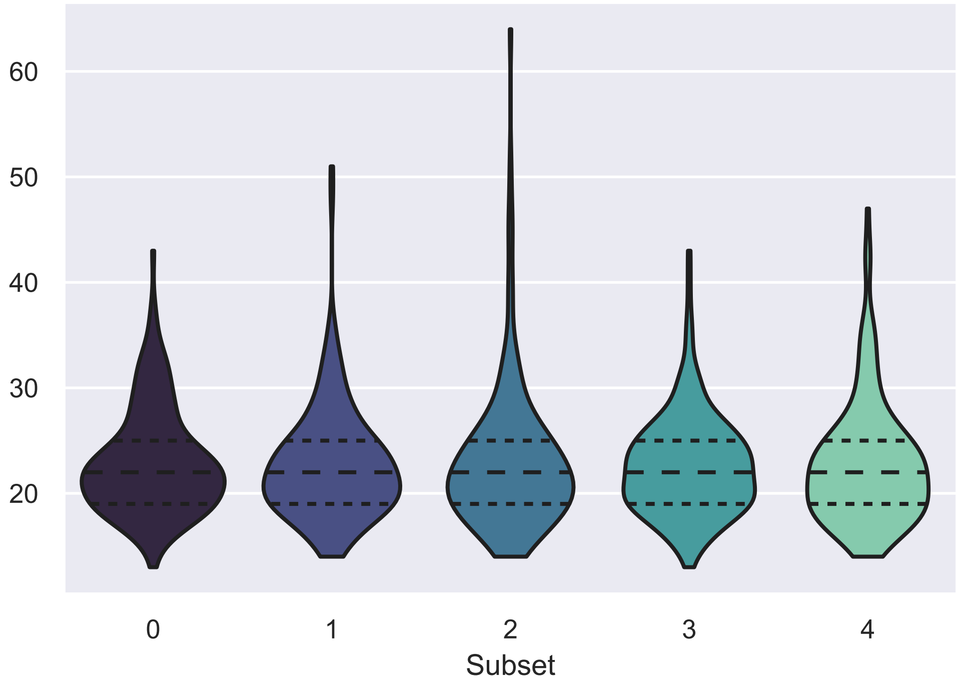
Feature	HP2 max	HP2 pos	L1	HP2 avg	R1	W1	K1	K2	I1	D1	S1	E1	I2	G1	C2	M2	R2	L2	G2	Q1	N1	Y2
Avg. Rank	1.0	2.0	3.2	4.0	5.8	8.6	12.6	14.0	14.6	15.6	17.2	17.6	18.8	18.8	20.2	21.4	21.6	21.6	21.6	22.0	22.6	25.0

V1	N2	D2	P1	Y1	H1	F1	V2	T1	F2	E2	T2	P2	M1	W2	C1	A2	S2	A1	H2	Q2
25.4	25.4	25.8	27.0	27.0	27.6	27.8	28.0	28.8	29.0	29.2	29.4	29.6	29.6	29.8	30.8	31.0	32.6	33.0	33.6	35.8

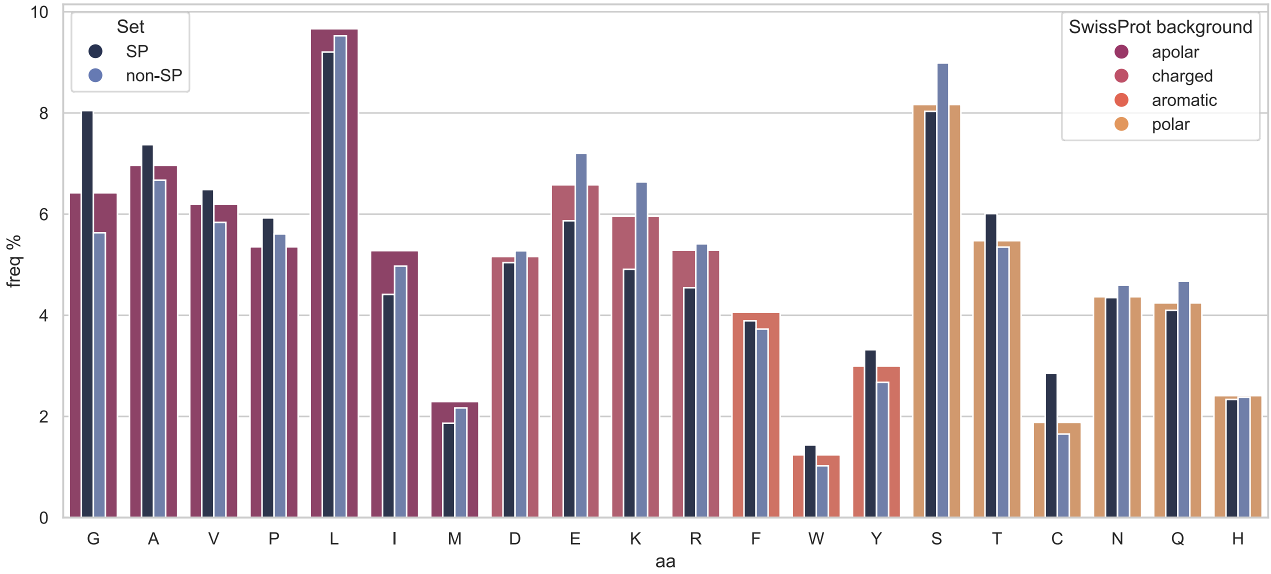
Supplementary Table S4. 5-fold CV scores and best hyperparameters for all SVM models and VH (average ± standard error)

Model	K	γ	C	MCC val	MCC test	ACC	Precision	Recall
SVM (N-COMP)	18	1	4	0.82 ± 0.01	0.79 ± 0.01	0.96 ± 0.00	0.84 ± 0.01	0.79 ± 0.01
SVM+C-COMP	19	scale	2	0.86 ± 0.01	0.83 ± 0.01	0.97 ± 0.00	0.88 ± 0.01	0.83 ± 0.01
SVM+HP	26	2	1	0.87 ± 0.01	0.84 ± 0.01	0.97 ± 0.00	0.86 ± 0.01	0.86 ± 0.01
SVM+HP2	22	4	2	0.89 ± 0.01	0.87 ± 0.01	0.97 ± 0.00	0.87 ± 0.00	0.89 ± 0.01
SVM+CH	22	4	4	0.82 ± 0.01	0.80 ± 0.02	0.96 ± 0.00	0.86 ± 0.01	0.79 ± 0.03
SVM+HP2+C-COMP	27	8	8	0.92 ± 0.01	0.89 ± 0.00	0.98 ± 0.00	0.91 ± 0.01	0.90 ± 0.01
SVM+HP2+HP	18	1	8	0.89 ± 0.01	0.88 ± 0.01	0.98 ± 0.00	0.88 ± 0.01	0.90 ± 0.01
SVM+HP2+CH	28	2	2	0.88 ± 0.01	0.87 ± 0.01	0.98 ± 0.00	0.88 ± 0.01	0.88 ± 0.02
SVM+HP2+C-COMP+HP	26	4	8	0.91 ± 0.00	0.89 ± 0.01	0.98 ± 0.00	0.89 ± 0.01	0.91 ± 0.01
SVM+HP2+C-COMP	27	8	8	0.92 ± 0.01	0.89 ± 0.00	0.98 ± 0.00	0.91 ± 0.01	0.90 ± 0.01
N-COMP+HP2+C-COMP+FF	27	8	8	0.89 ± 0.01	**0.89 ± 0.01**	0.98 ± 0.00	0.90 ± 0.01	0.91 ± 0.02
	Score threshold							
VH	9.20 ± 0.23			/	**0.69 ± 0.01**	0.94 ± 0.00	0.73 ± 0.03	0.72 ± 0.03

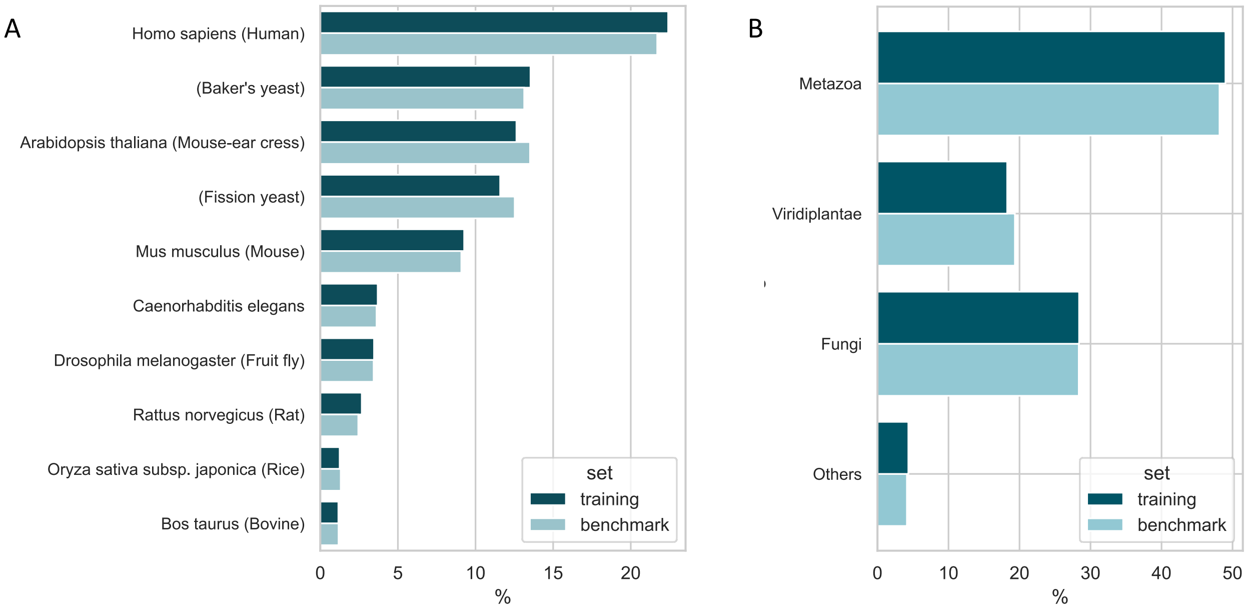
Figures



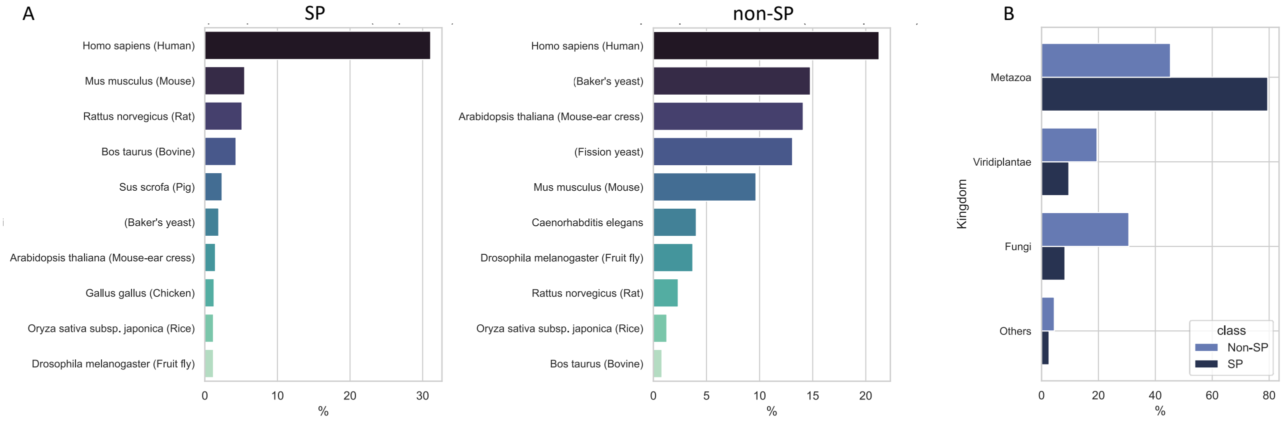
Supplementary Figure S1. SP length distributions after splitting training set in 5 subsets.



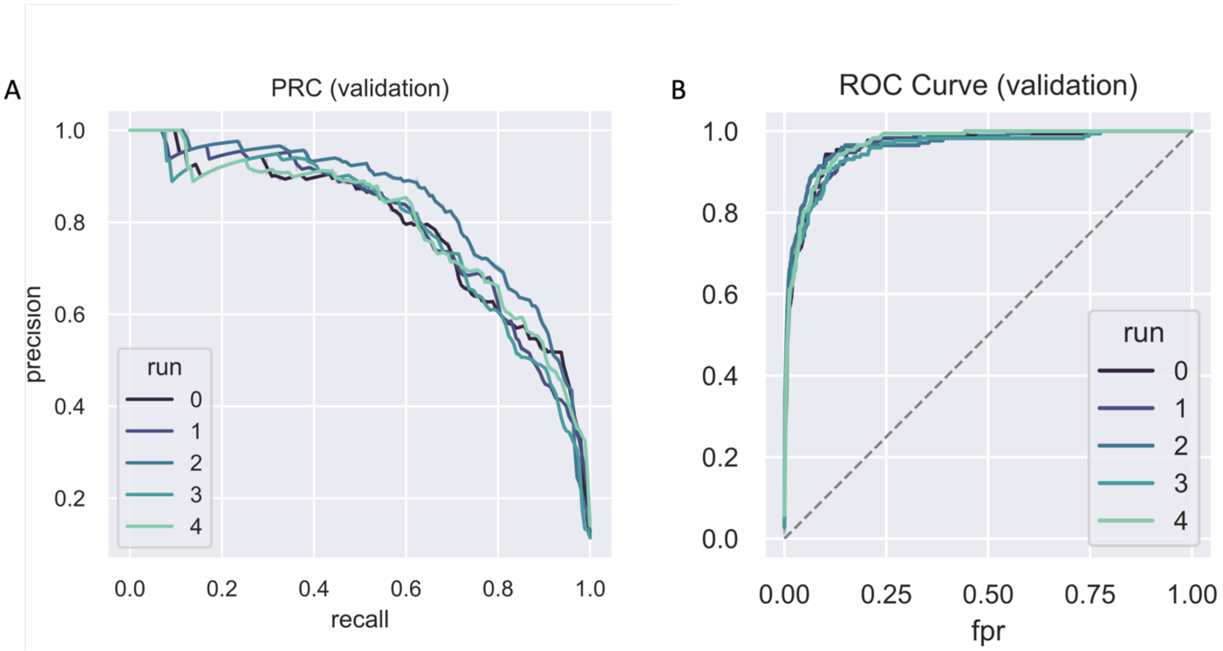
Supplementary Figure S2. Global residue composition (SP vs non-SP).



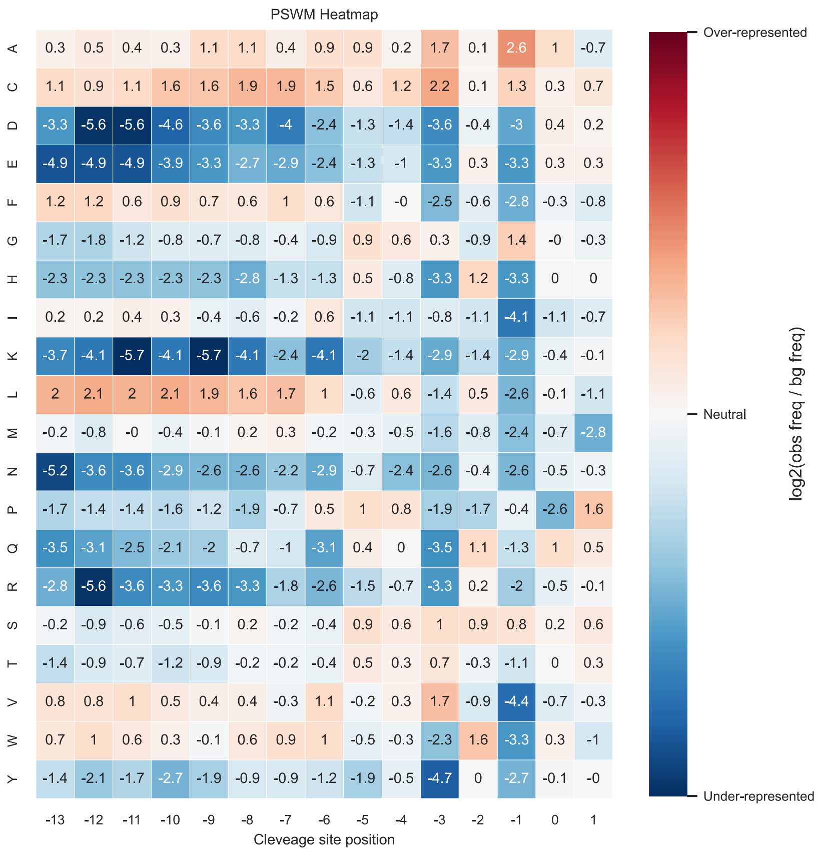
Supplementary Figure S3. A. Top-10 species % distributions by set. B. Kingdom % distribution by set.



Supplementary Figure S4. A. Top-10 species % distributions by class . B. Kingdom % distribution by class.

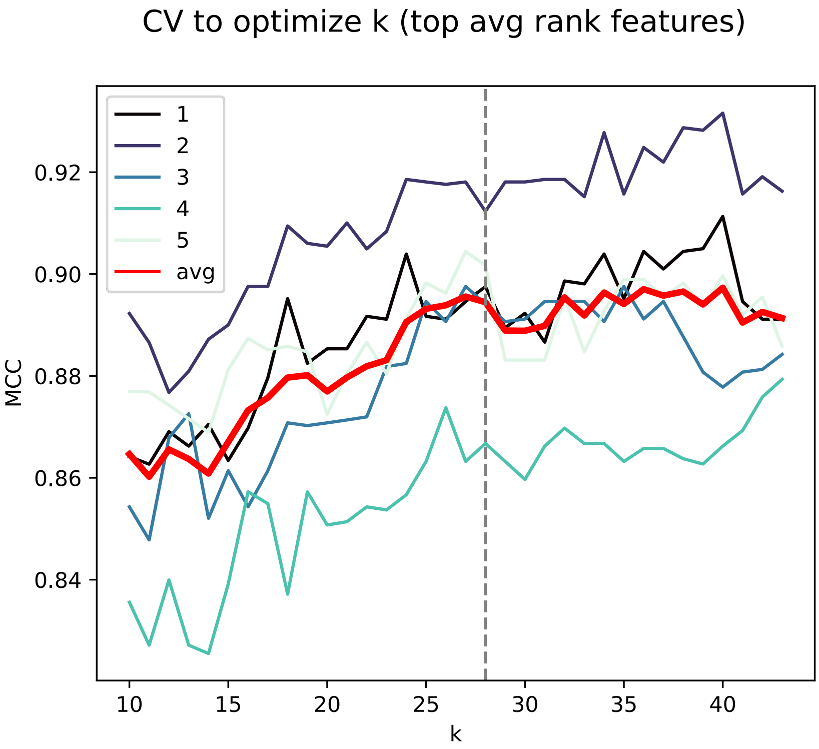


Supplementary Figure S5. A. Precision recall curve (PRC) generated in VH 5-fold CV. B. Receiver operating characteristic (ROC) curve generated in VH 5-fold CV.

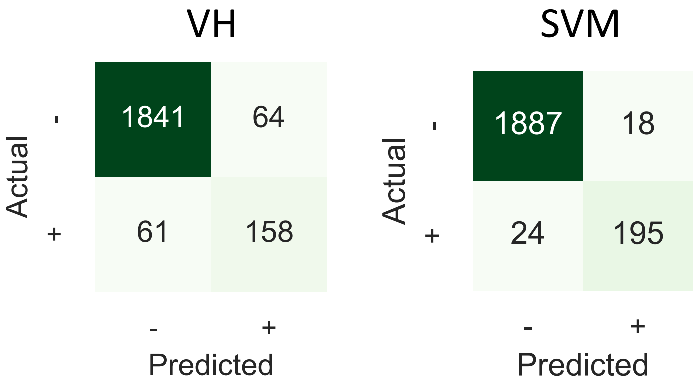


	-13	-12	-11	-10	-9	-8	-7	-6	-5	-4	-3	-2	-1	0	1
A	0.3	0.5	0.4	0.3	1.1	1.1	0.4	0.9	0.9	0.2	1.7	0.1	2.6	1	-0.7
C	1.1	0.9	1.1	1.6	1.6	1.9	1.9	1.5	0.6	1.2	2.2	0.1	1.3	0.3	0.7
D	-3.3	-5.6	-5.6	-4.6	-3.6	-3.3	-4	-2.4	-1.3	-1.4	-3.6	-0.4	-3	0.4	0.2
E	-4.9	-4.9	-4.9	-3.9	-3.3	-2.7	-2.9	-2.4	-1.3	-1	-3.3	0.3	-3.3	0.3	0.3
F	1.2	1.2	0.6	0.9	0.7	0.6	1	0.6	-1.1	-0	-2.5	-0.6	-2.8	-0.3	-0.8
G	-1.7	-1.8	-1.2	-0.8	-0.7	-0.8	-0.4	-0.9	0.9	0.6	0.3	-0.9	1.4	-0	-0.3
H	-2.3	-2.3	-2.3	-2.3	-2.3	-2.8	-1.3	-1.3	0.5	-0.8	-3.3	1.2	-3.3	0	0
I	0.2	0.2	0.4	0.3	-0.4	-0.6	-0.2	0.6	-1.1	-1.1	-0.8	-1.1	-4.1	-1.1	-0.7
K	-3.7	-4.1	-5.7	-4.1	-5.7	-4.1	-2.4	-4.1	-2	-1.4	-2.9	-1.4	-2.9	-0.4	-0.1
L	2	2.1	2	2.1	1.9	1.6	1.7	1	-0.6	0.6	-1.4	0.5	-2.6	-0.1	-1.1
M	-0.2	-0.8	-0	-0.4	-0.1	0.2	0.3	-0.2	-0.3	-0.5	-1.6	-0.8	-2.4	-0.7	-2.8
N	-5.2	-3.6	-3.6	-2.9	-2.6	-2.6	-2.2	-2.9	-0.7	-2.4	-2.6	-0.4	-2.6	-0.5	-0.3
P	-1.7	-1.4	-1.4	-1.6	-1.2	-1.9	-0.7	0.5	1	0.8	-1.9	-1.7	-0.4	-2.6	1.6
Q	-3.5	-3.1	-2.5	-2.1	-2	-0.7	-1	-3.1	0.4	0	-3.5	1.1	-1.3	1	0.5
R	-2.8	-5.6	-3.6	-3.3	-3.6	-3.3	-1.8	-2.6	-1.5	-0.7	-3.3	0.2	-2	-0.5	-0.1
S	-0.2	-0.9	-0.6	-0.5	-0.1	0.2	-0.2	-0.4	0.9	0.6	1	0.9	0.8	0.2	0.6
T	-1.4	-0.9	-0.7	-1.2	-0.9	-0.2	-0.2	-0.4	0.5	0.3	0.7	-0.3	-1.1	0	0.3
V	0.8	0.8	1	0.5	0.4	0.4	-0.3	1.1	-0.2	0.3	1.7	-0.9	-4.4	-0.7	-0.3
W	0.7	1	0.6	0.3	-0.1	0.6	0.9	1	-0.5	-0.3	-2.3	1.6	-3.3	0.3	-1
Y	-1.4	-2.1	-1.7	-2.7	-1.9	-0.9	-0.9	-1.2	-1.9	-0.5	-4.7	0	-2.7	-0.1	-0

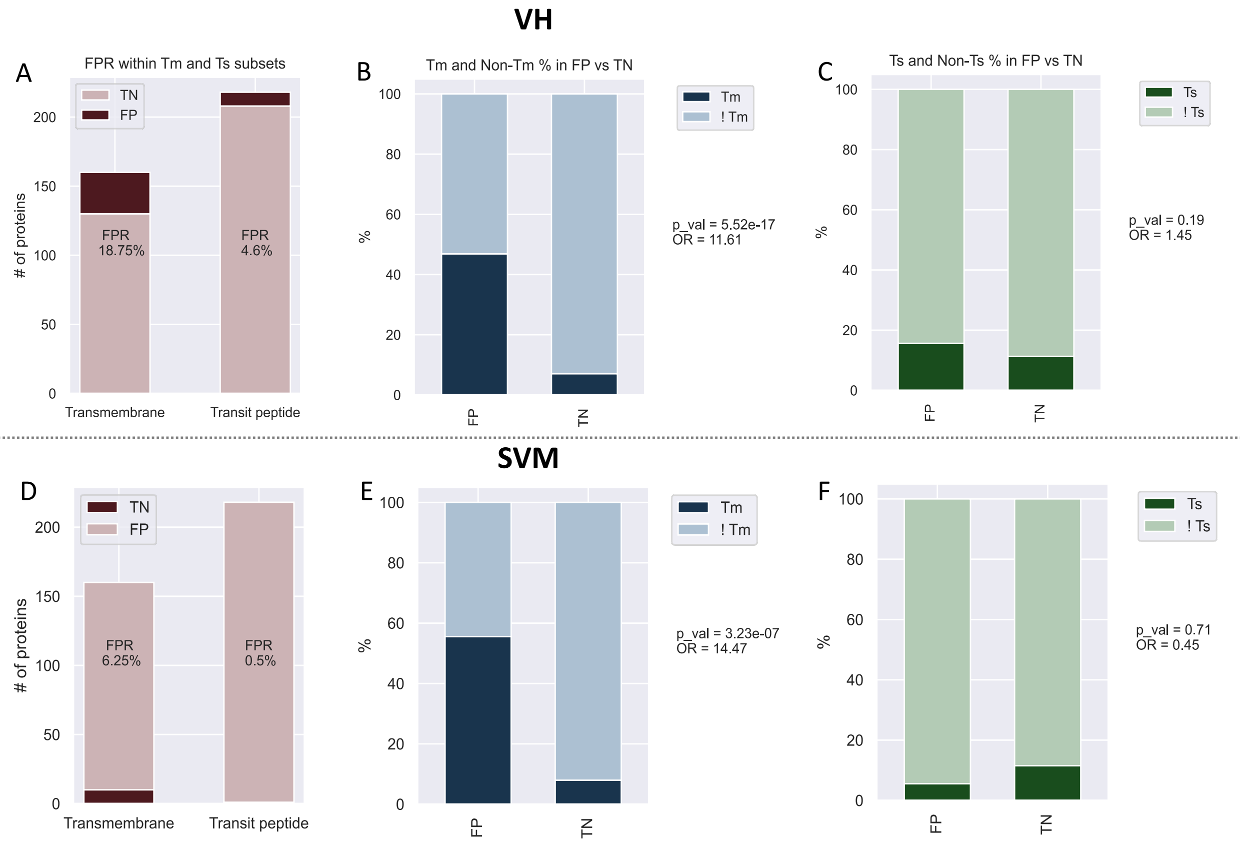
Supplementary Figure S6. VH PSWM computed with the whole training set.



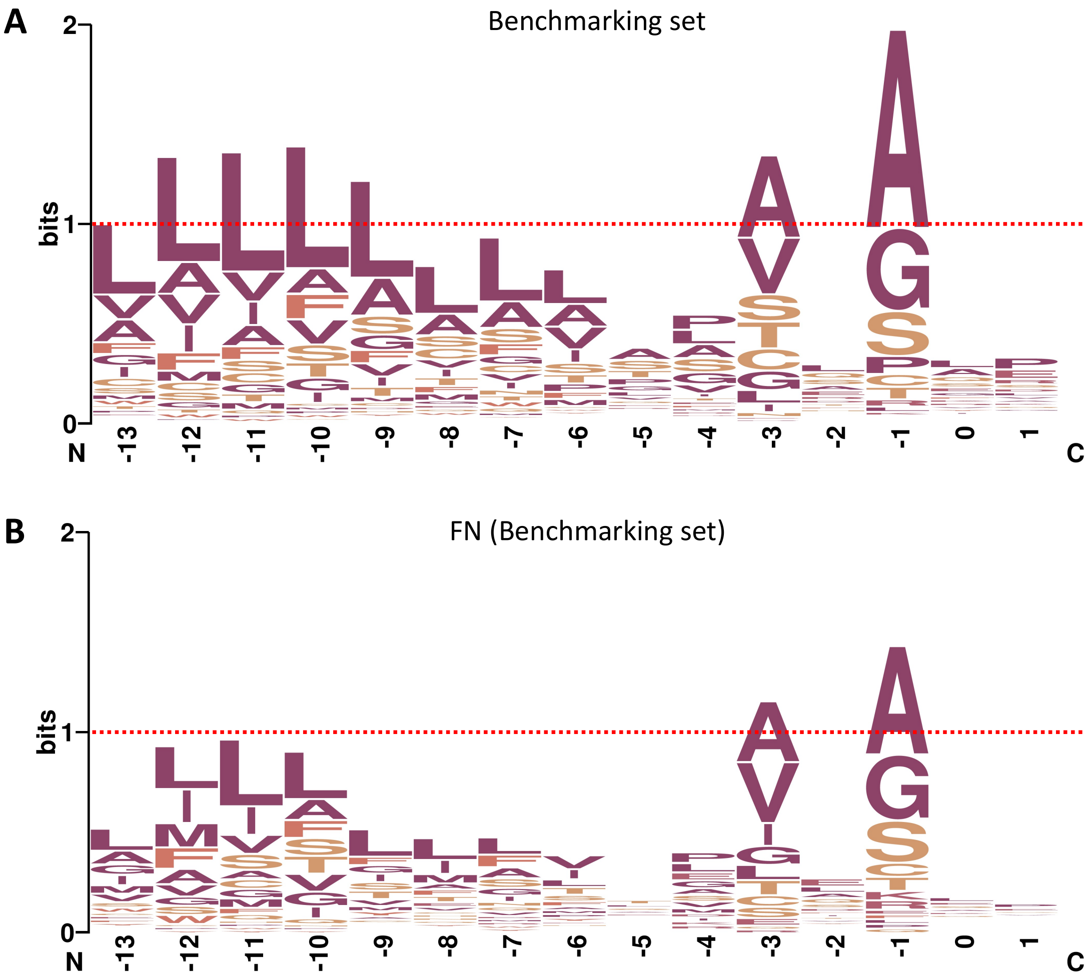
Supplementary Figure S7. Permutation feature importance (PFI) CV performance (MCC) trend as function of k, where k is the number of top features selected in the feature filtering.



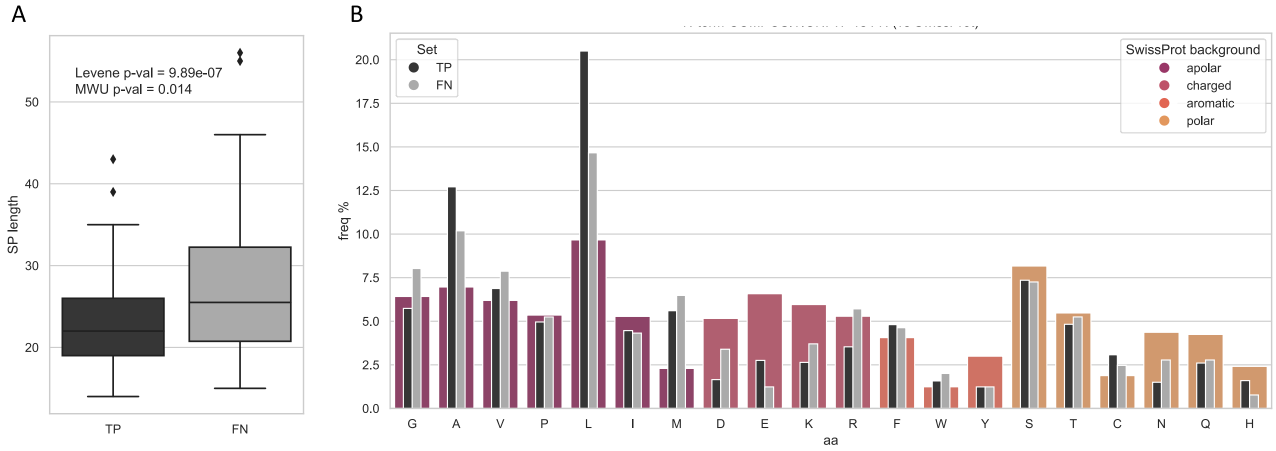
Supplementary Figure S8. Confusion matrices generated in benchmarking VH and SVM.



Supplementary Figure S9. False positive analyses of both VH and SVM. Comparison of FPR_{Tm} with FPR_{Ts} (A and D). Comparison of transmembrane proteins proportions between TN and FP (Fisher's exact)(B and E). Comparison of transit peptide proteins proportions between TN and FP (Fisher's exact)(C and F)



Supplementary Figure S10. A. Cleavage site (CS) sequence logo (-13,+2) computed on SP sequences (in the whole benchmarking set) **B.** CS sequence logo (-13,+2) computed only on FN resulted from VH benchmarking.



Supplementary Figure S11. A. The SP length (CS position) distributions are compared between FN and TP resulted from SVM benchmarking. **B.** N-terminus residue compositions (until position K = 28) are compared between FN and TP resulted from SVM benchmarking.