*Supplementary material*

# Evaluating Signal Peptide Prediction in Eukaryotes: support vector machine VS position-weight matrix approach

Irene D'Onofrio[1, *]

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

*To whom correspondence should be addressed.

**Supplementary Table S1**. The entire dataset is in dataset.tsv

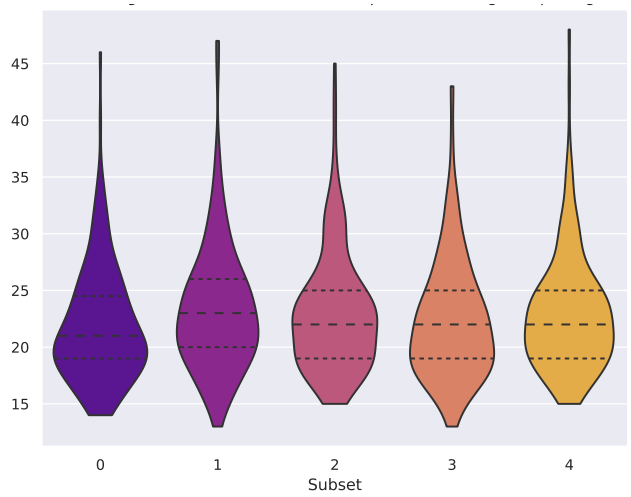
**Supplementary Table S2.** Model hyperparameters optimized with grid search.

| Model | Hyperparameter | Values tested |
|---|---|---|
| SVM | C | 1,2,4,8 |
| SVM | gamma | 1, 2, 'scale' (i.e. gamma = 1 / (n_features * X.var())) |
| SVM | K | 18,19,20,21,22,23,24,25,26 |

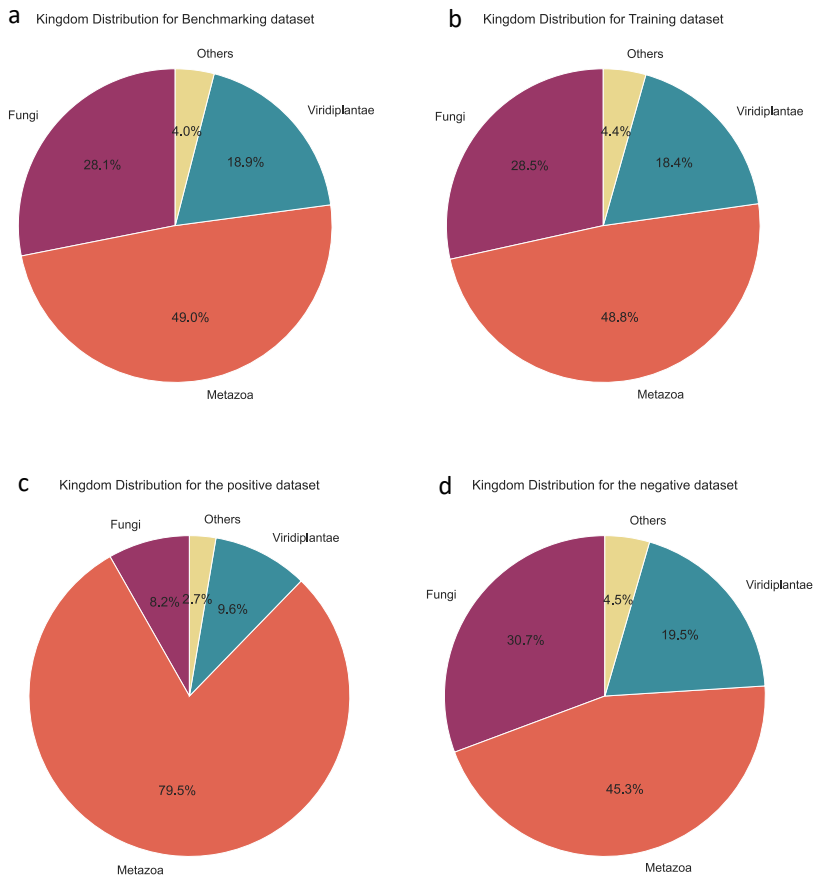
**Supplementary Table S3**. Cross validation results for von Heijne and each SVM model tested.

| model | Optimal threshold | | | ACC | MCC | Precision | Recall | F1-score |
|---|---|---|---|---|---|---|---|---|
| von Heijne | 9.06 ± 0.31 | | | 0.94 ± 0.00 | 0.67 ± 0.01 | 0.70 ± 0.04 | 0.72 ± 0.04 | 0.70 ± 0.01 |
| **model** | **C** | **gamma** | **K** | **ACC** | **MCC** | **Precision** | **Recall** | **F1-score** |
| SVM-Ncomp | 4 | scale | 20 | 0.96 ¬± 0.00 | 0.77 ¬± 0.02 | 0.83 ¬± 0.02* | 0.75 ± 0.03 | 0.78 ± 0.03 |
| SVM-Ncomp-gcomp | 2 | scale | 19 | 0.97 ¬± 0.00 | 0.84 ¬± 0.02 | 0.89 ¬± 0.01 | 0.82 ¬± 0.02 | 0.85 ± 0.02 |
| SVM-Ncomp-hp | 8 | 2 | 19 | 0.97 ¬± 0.00 | 0.83 ¬± 0.02 | 0.86 ¬± 0.01 | 0.83 ¬± 0.03 | 0.84 ± 0.02 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| SVM-Ncomp-ghp | 8 | scale | 18 | 0.96 ¬± 0.00 | 0.78 ¬± 0.02 | 0.85 ¬± 0.01 | 0.77 ¬± 0.03 | 0.80 ± 0.02 |
| SVM-Ncomp-ch | 8 | scale | 20 | 0.96 ¬± 0.00 | 0.78 ¬± 0.02 | 0.83 ¬± 0.02 | 0.78 ¬± 0.02 | 0.80 ± 0.02 |
| SVM-comp-ht | 8 | 1 | 19 | 0.96 ¬± 0.00 | 0.78 ¬± 0.02 | 0.83 ¬± 0.01 | 0.78 ¬± 0.02 | 0.81 ± 0.02 |
| SVM-Ncomp-tmt | 8 | 2 | 20 | 0.97 ¬± 0.00 | 0.85 ¬± 0.01 | 0.87 ¬± 0.01 | 0.86 ¬± 0.01 | 0.86 ¬± 0.01 |
| | | | | | | | | |
| SVM-Ncomp-tmt-gcomp | 8 | 2 | 26 | 0.98 ¬± 0.00 | 0.87 ¬± 0.01 | 0.90 ¬± 0.01 | 0.87 ¬± 0.02 | 0.89 ¬± 0.01 |
| SVM-Ncomp-tmt-hp | 8 | 2 | 26 | 0.97 ¬± 0.00 | 0.85 ¬± 0.01 | 0.86 ¬± 0.01 | 0.87 ¬± 0.02 | 0.86 ¬± 0.01 |
| SVM-Ncomp-tmt-ghp | 8 | 2 | 21 | 0.97 ¬± 0.00 | 0.85 ¬± 0.01 | 0.88 ¬± 0.01 | 0.86 ¬± 0.01 | 0.87 ¬± 0.01 |
| SVM-Ncomp-tmt-ch | 4 | 2 | 23 | 0.97 ¬± 0.00 | 0.86 ¬± 0.01 | 0.87 ¬± 0.01 | 0.87 ¬± 0.01 | 0.87 ¬± 0.01 |
| SVM-Ncomp-tmt-ht | 8 | 2 | 23 | 0.97 ¬± 0.00 | 0.85 ¬± 0.01 | 0.86 ¬± 0.00 | 0.86 ¬± 0.01 | 0.86 ¬± 0.01 |
| | | | | | | | | |
| SVM-Ncomp-tmt-gcomp-hp | 8 | 2 | 23 | 0.98 ¬± 0.00 | 0.89 ¬± 0.00 | 0.91 ¬± 0.01 | 0.89 ¬± 0.00 | 0.90 ¬± 0.00 |
| SVM-Ncomp-tmt-gcomp-ghp | 8 | 2 | 22 | 0.98 ¬± 0.00 | 0.89 ¬± 0.00 | 0.91 ¬± 0.01 | 0.89 ¬± 0.01 | 0.90 ¬± 0.00 |
| SVM-Ncomp-tmt-gcomp-ch | 2 | 2 | 23 | 0.98 ¬± 0.00 | 0.88 ¬± 0.01 | 0.90 ¬± 0.01 | 0.89 ¬± 0.01 | 0.89 ¬± 0.01 |
| SVM-Ncomp-tmt-gcomp-ht | 8 | 2 | 22 | 0.98 ¬± 0.00 | 0.88 ¬± 0.01 | 0.91 ¬± 0.00 | 0.88 ¬± 0.01 | 0.89 ¬± 0.00 |

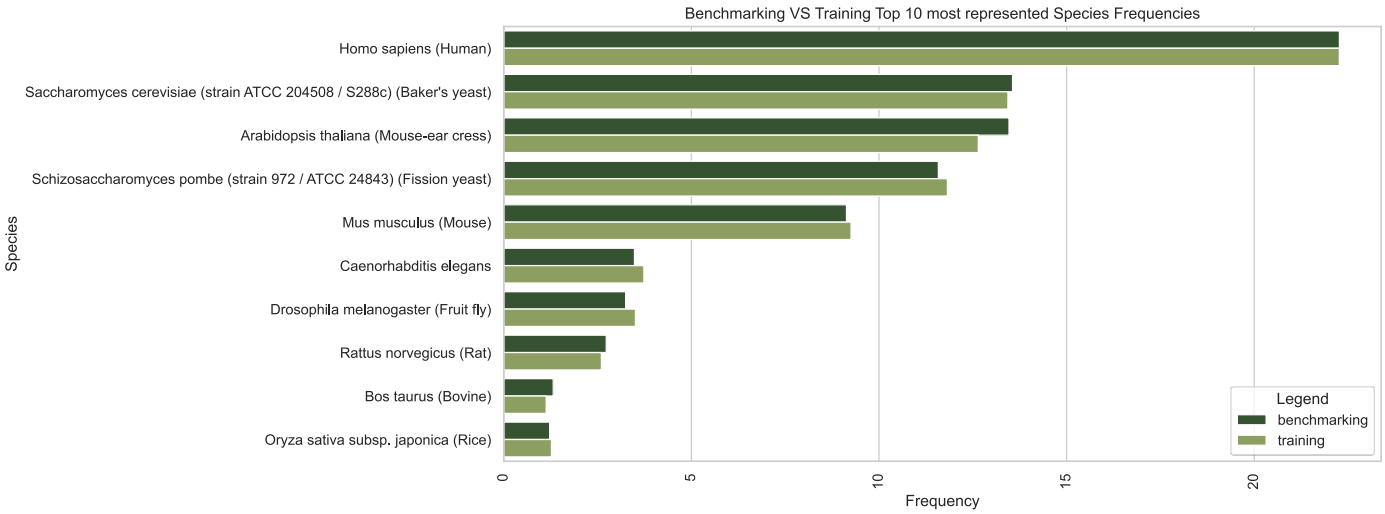



**Supplementary Figure S1.** Signal Peptide length distribution over the 5 subsets; the homogeneity of the distribution is an indication of a fair splitting.

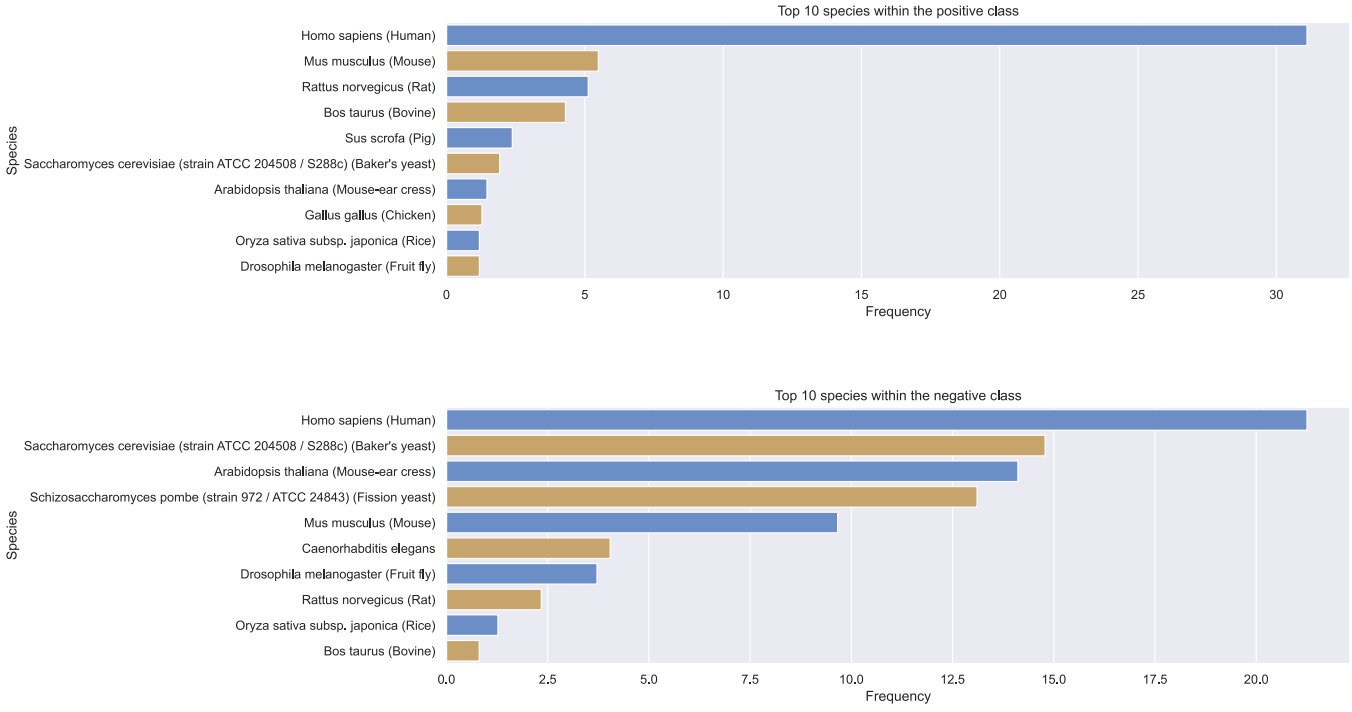



**Supplementary Figure S2.**

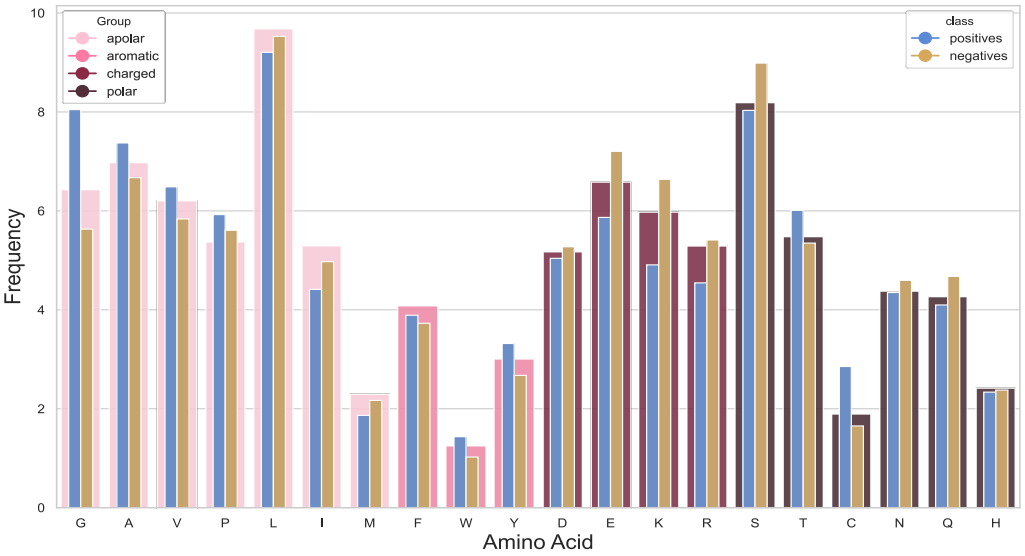
Distribution of the kingdoms within Benchmarking (a) and Training (b) datasets and SP (c) and non-SP proteins (d).




**Supplementary Figure S3**. Frequency of the top 10 most represented species within training and benchmarking dataset

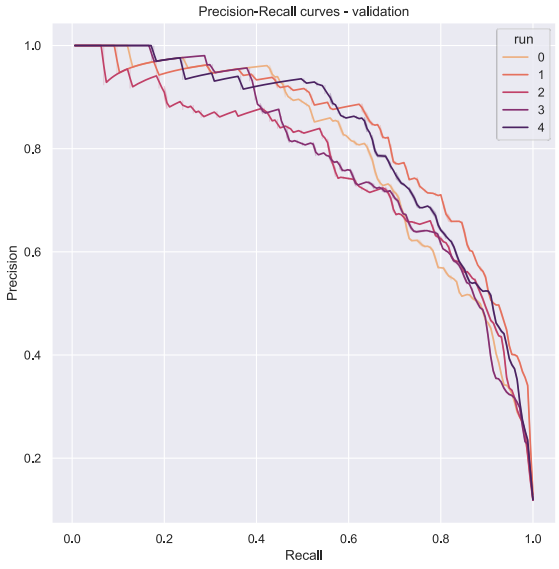



**Supplementary Figure S4**. Frequency of the top 10 most represented species among the SP and non-SP proteins.

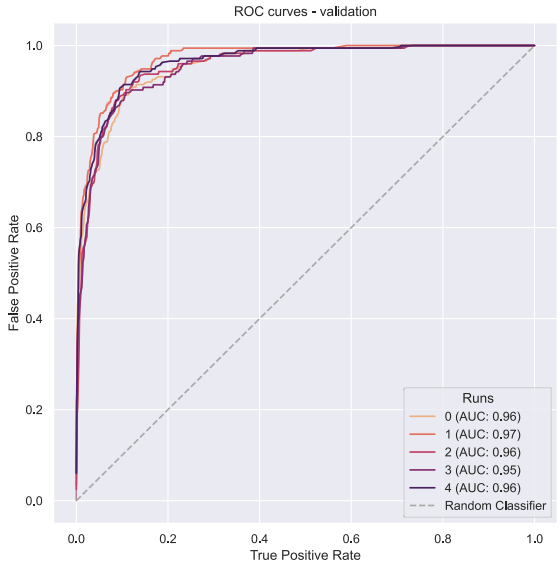



**Supplementary Figure S5.** Residue composition over all the protein length in SP proteins and non-SP proteins.

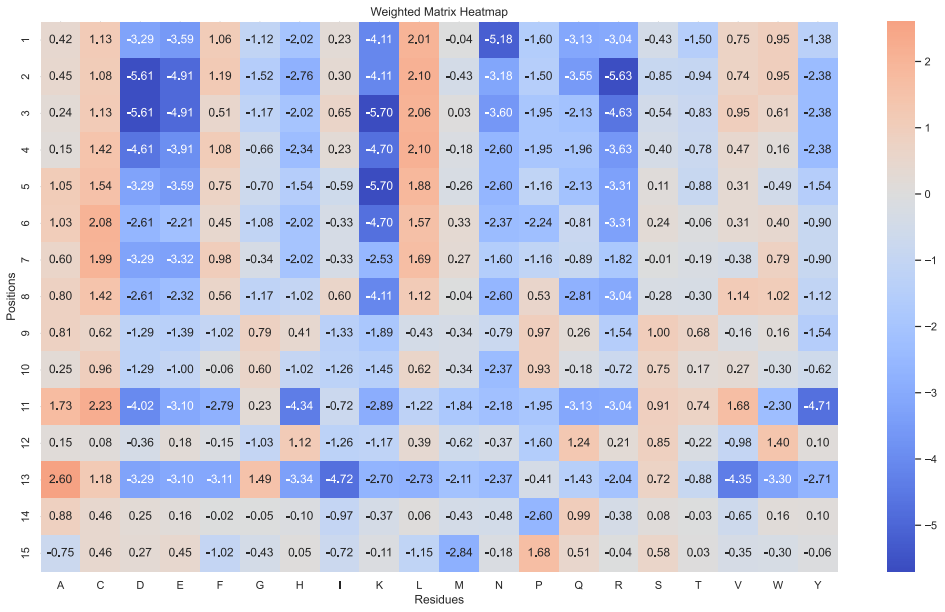
a



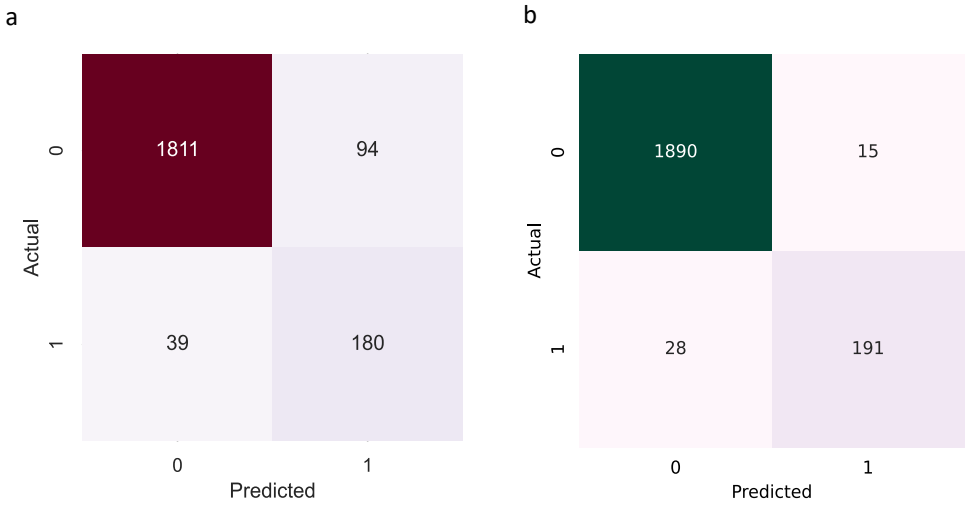

b



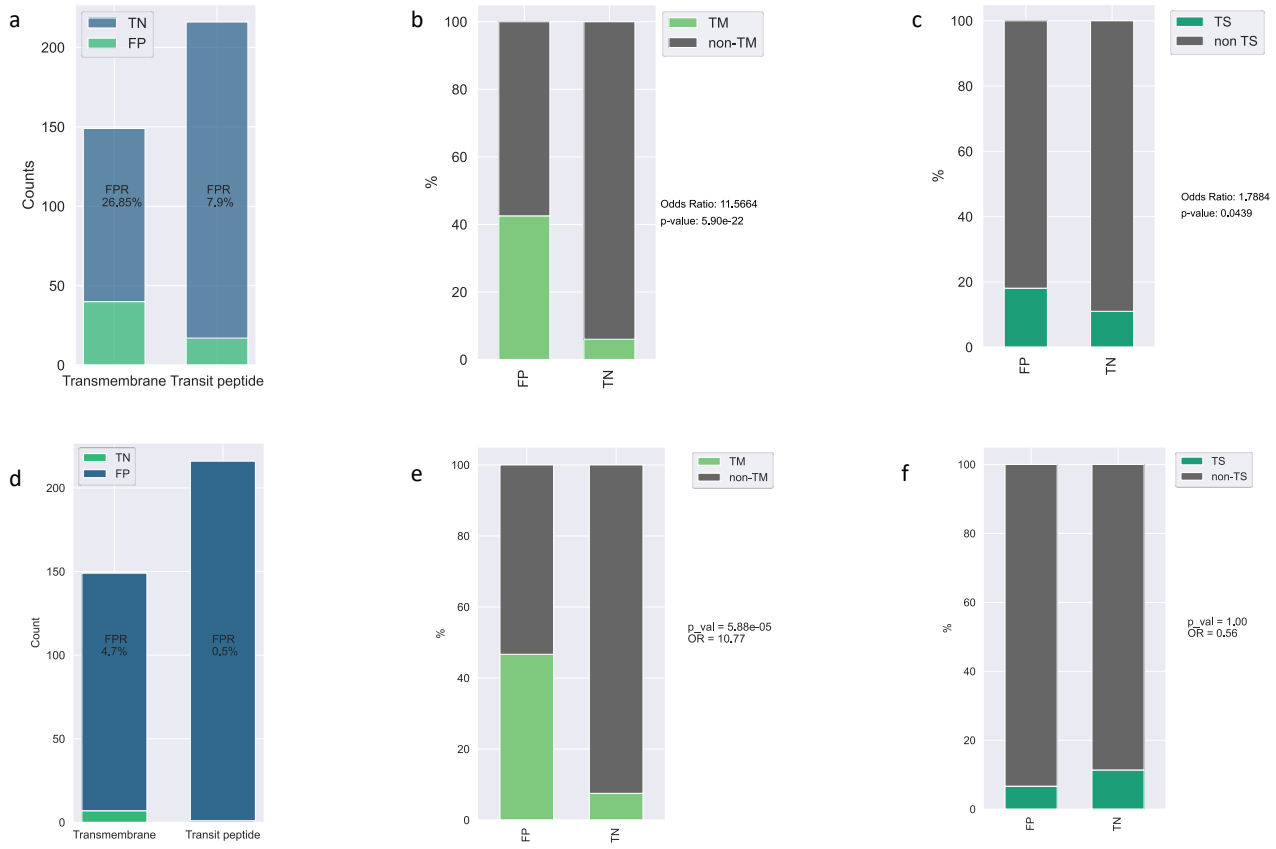

**Supplementary Figure 6**. **a.** Precision-Recall curves of the Von Heijne method cross-validation procedure. **b.** ROC curves of the Von Heijne method cross-validation procedure.




**Supplementary Figure S7.** Position Specific Scoring Matrix computed on the entire training dataset.

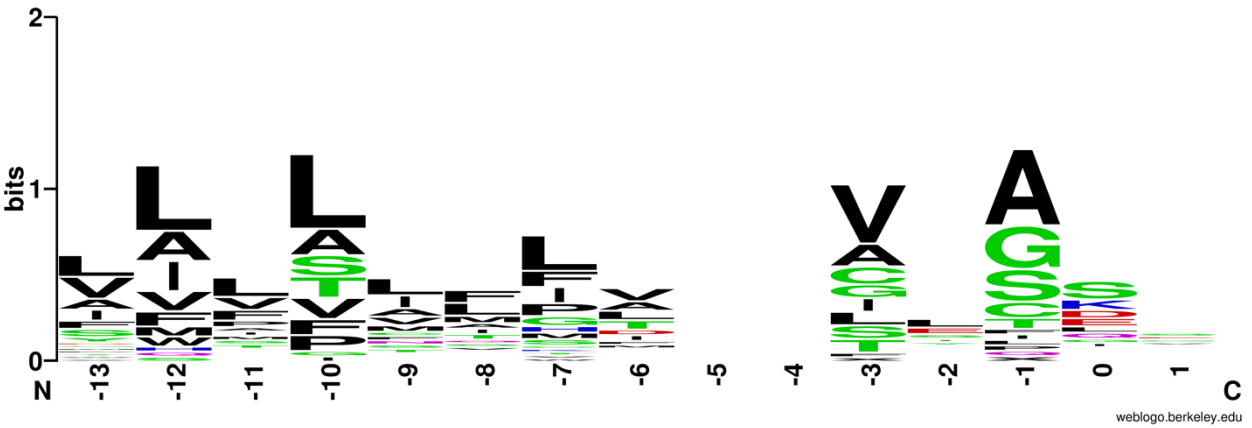



**Supplementary Figure S8.** Confusion matrices of the von Heijne method (a) and the SVM-Ncomp-tmt-gcomp-ghp' model.

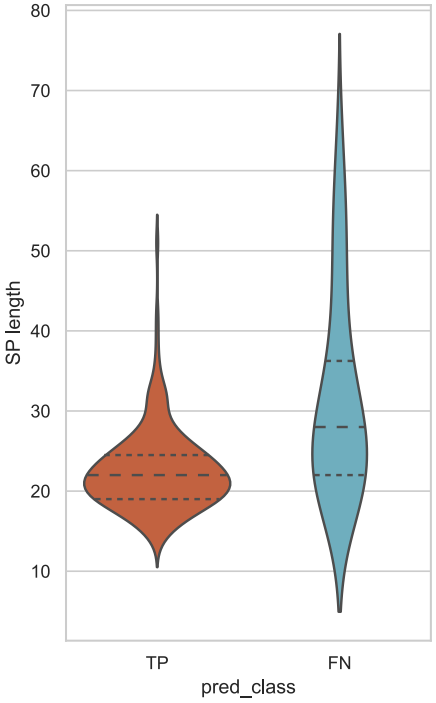



.

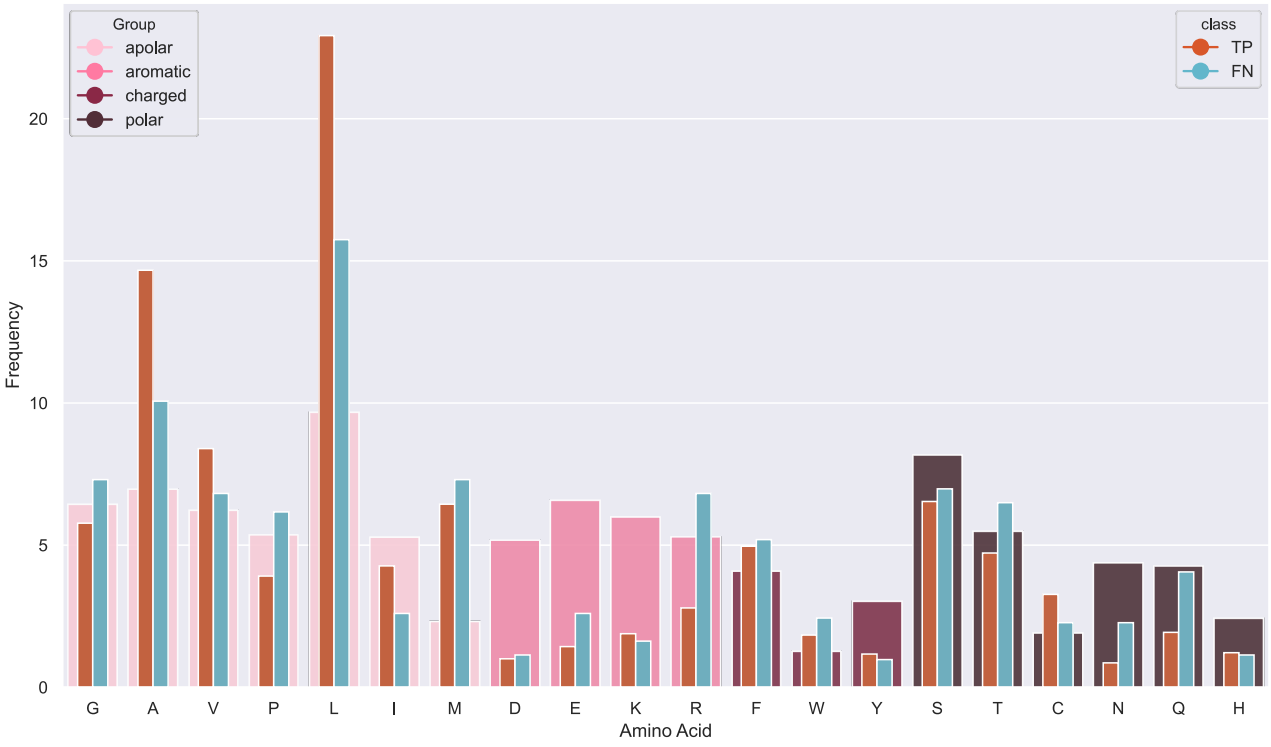
**Supplementary Figure S9.** False positive analyses. FPR rates for both TM and TS in von Heijne (a) and SVM (d) models. In proportion, the enrichment of TM within the false positive predictions is significative for both von Heijne (b) and SVM model (e), with Fisher-exact P of 5.90e-22 and 5.88e-5 respectively. As for the TS, only in von Heijne the p-value is significant (c), contrary to the SVM model (f).




**Supplementary Figure S10.** Sequence logo of the region -13 +2 of the sequences wrongly predicted as negatives (FN). The information content is clearly lower over all the length analyzed. Indeed, from position -13 to -6, the typical abundance of leucine, alanine and valine is not evident. The same can be stated for the positions -3 -1.




**Supplementary Figure S11.** SP length distribution within the true positive predictions and false negative predictions. In the FN the distribution is asymmetric, with the minimum and maximum lengths extending to more extreme values. Also, the mean is clearly far from the mean individuated for the SP (~23)




**Supplementary Figure S12.** Residue composition of the SP sequences within true positive predictions and false negative predictions.