Supplementary materials

Here are collected all graphical representations that didn't make in the final version of the report.

From fig. 1 to fig. 3 we display pictures related to the data analysis subsection of the Methods section.

From fig. 4 to fig. 7, we report pictures related to the benchmark evaluation subsection (false positive and false negative) of the Results section.

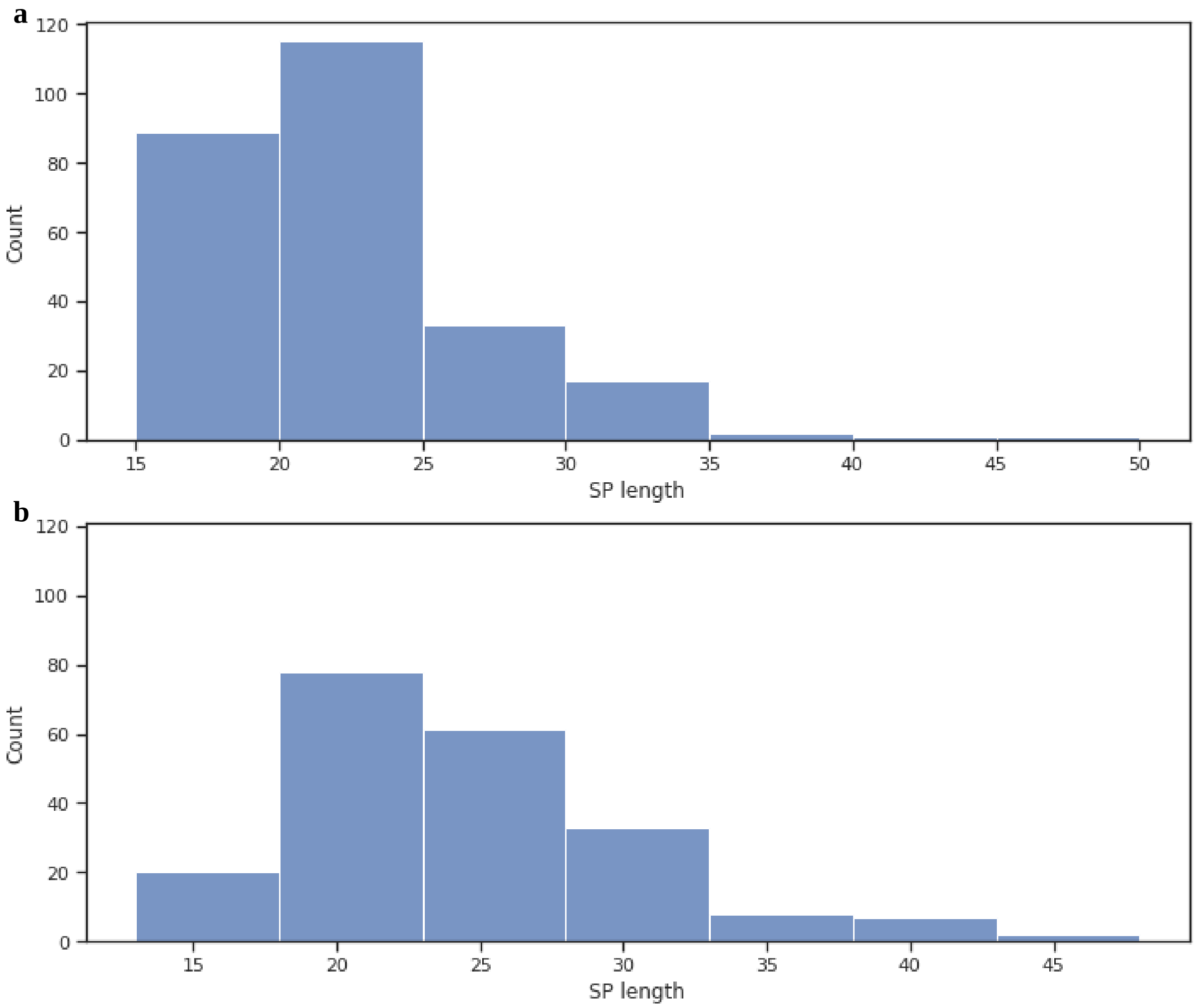

Fig. 1. Signal peptide length distribution. Panel **a** display the distribution in the training dataset; panel **b** in the benchmark dataset. The trend of both distribution is similar, reflecting indeed the average signal peptide length, which ranges between 20/25 residues.

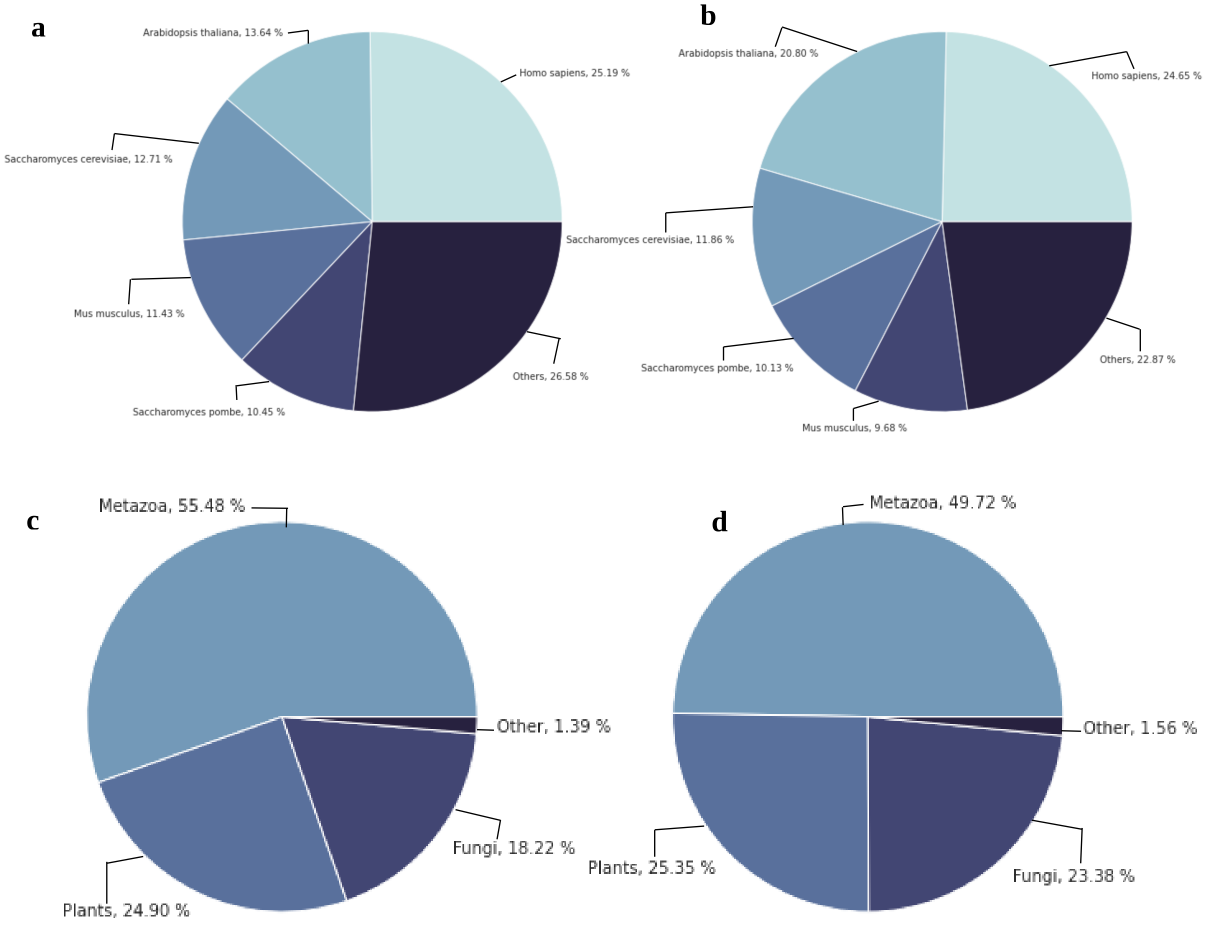

Fig. 2. Taxa and kingdom distribution. Upper panels represent the taxa distribution in the training dataset (**a**) and benchmark dataset (**b**), focusing on the top 5 most populated ones (*Homo sapiens* at the top in both datasets). Lower panels display, in the same order (**c** for training, **d** for benchmark), the kingdom distribution (Metazoa at the top). We can observe that we have at least one representative taxa for each kingdom displayed.

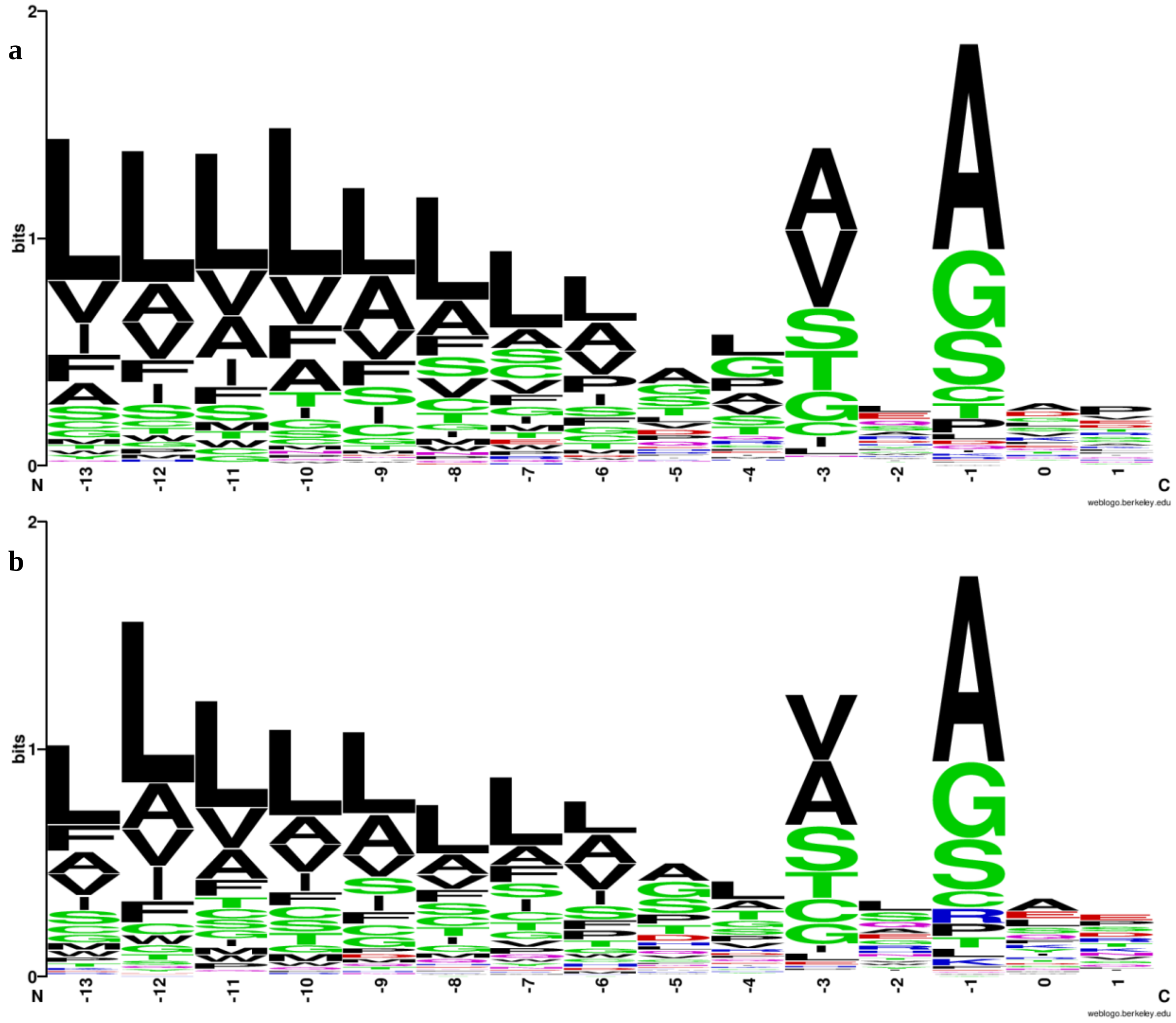

Fig. 3. Consensus sequence logo. Panel **a** displays the consensus sequence logo of the training dataset. Panel **b** the consensus sequence logo of the benchmark dataset. Both consensus sequences depict signal peptide composition from position -13 up to +2 from the cleavage site (in the logo, position 0 and 1 corresponds respectively to position +1 and +2 of the signal peptide). Relevant subregions of signal peptides are the hydrophobic core (-13 to -6) and the motif directly preceding the cleavage site (-3 to +1), which highly-hydrophobic character represents one of the most compelling property of this type of sorting signals. The following color scheme is followed: polar residues (G, S, T, Y, C, Q, N) in green; basic residues (K, R, H) in blue; acidic residues (D, E) in red; hydrophobic residues (A, V, L, I, P, W, F, M) in black.

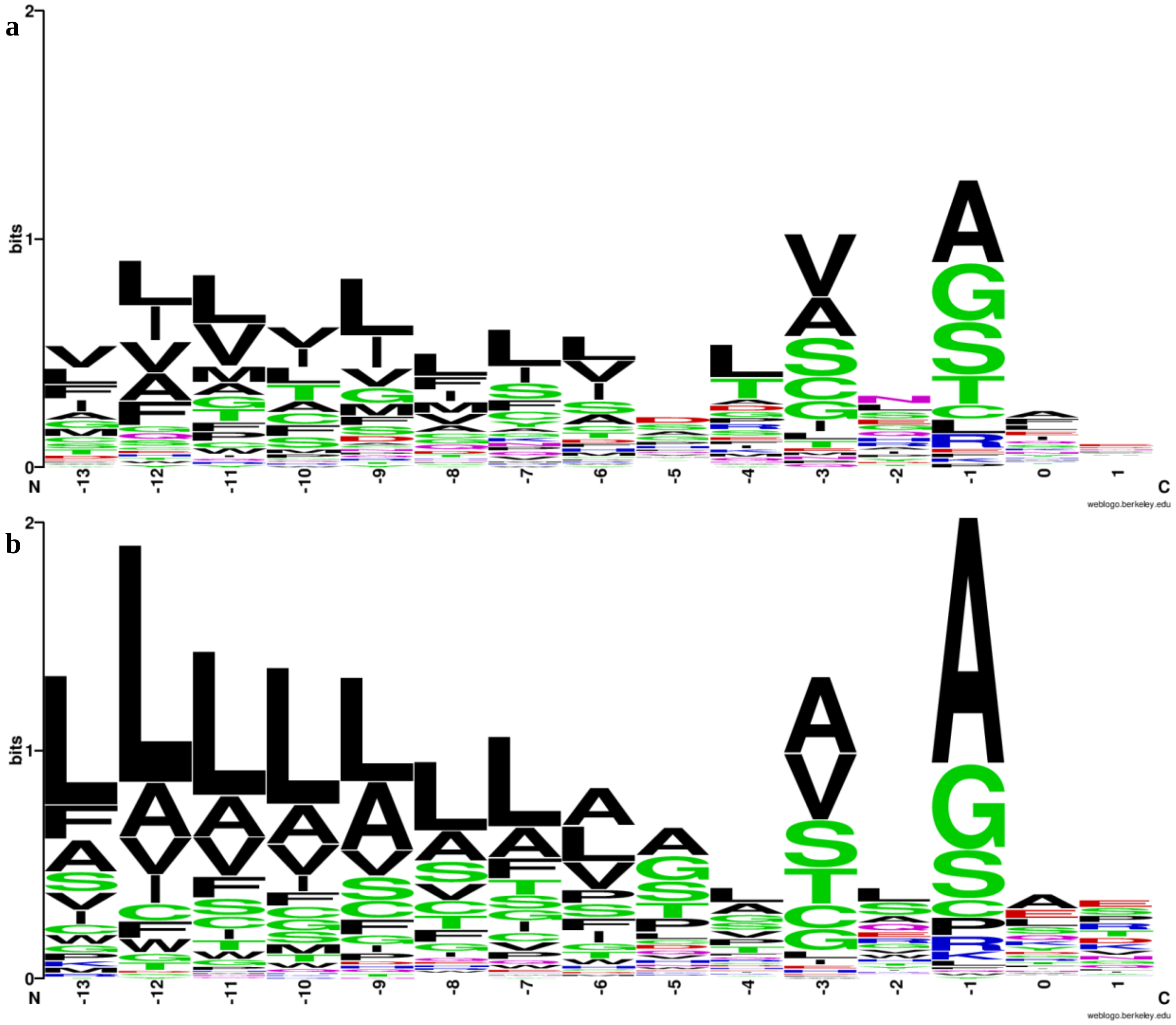

Fig. 4. von Heijne false negative vs true positive consensus sequence logo. Panel **a** represent the consensus sequence of false negatives; panel **b** the consensus sequence of true positives. Both consensus sequences depict signal peptide composition from position -13 up to +2 from the cleavage site (in the logo, position 0 and 1 corresponds respectively to position +1 and +2 of the signal peptide). We can see that, in the false negatives, the most representative hydrophobic residues Leu (L) and Ala (A) are not so dominant anymore (their frequency is now comparable to that of other residues, both hydrophobic and hydrophilic). The following color scheme is followed: polar residues (G, S, T, Y, C, Q, N) in green; basic residues (K, R, H) in blue; acidic residues (D, E) in red; hydrophobic residues (A, V, L, I, P, W, F, M) in black.

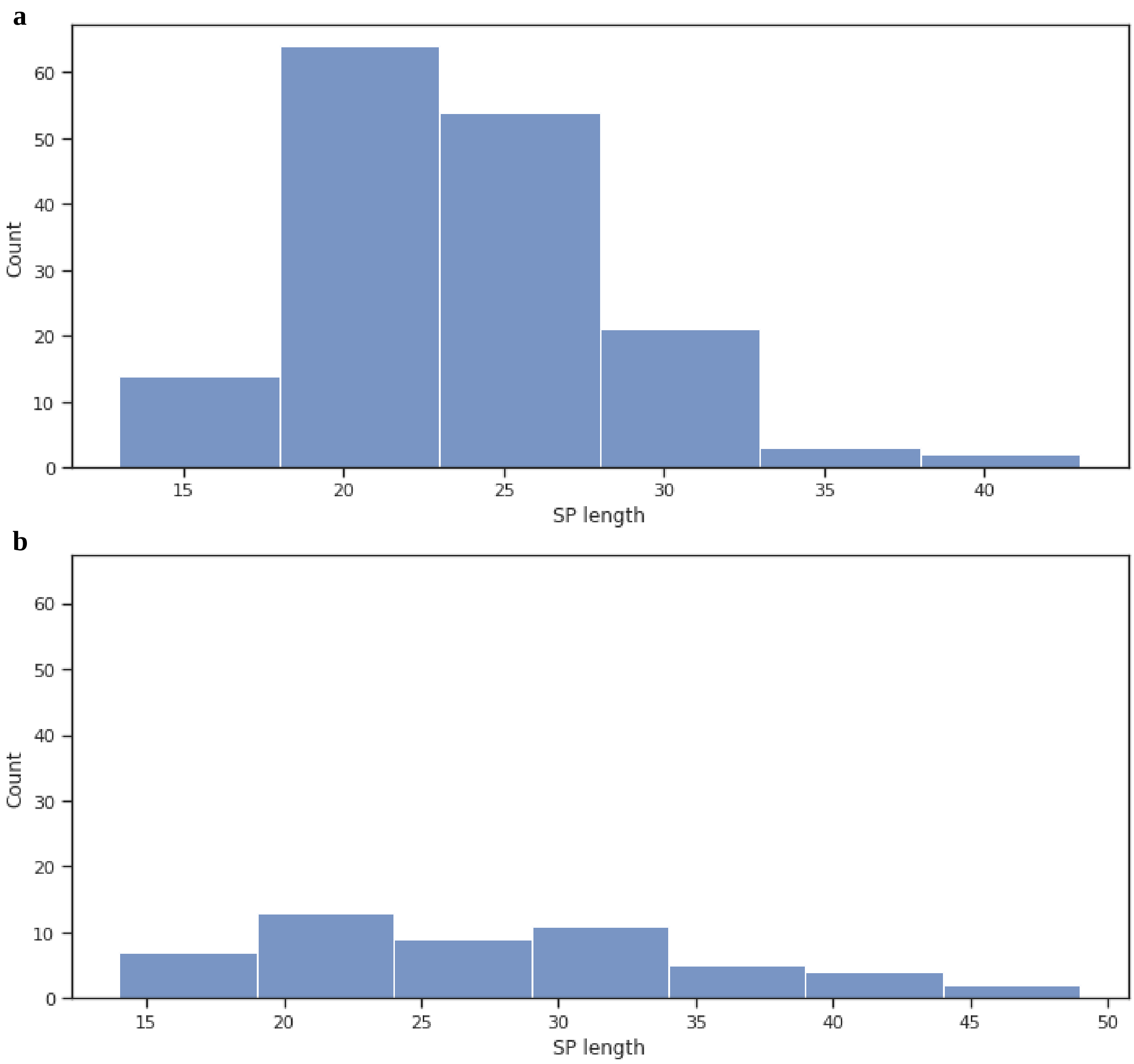

Fig. 5. Support vector machine signal peptide length distribution. Panel **a** represents signal peptides length distribution in the true positive set, panel **b** in the false negative one. Compared to the set of true signal peptides, false negatives trend is more uniformly distributed, in which the average signal peptide length is 27-res long vs the commonly found signal peptide of length 23-res.

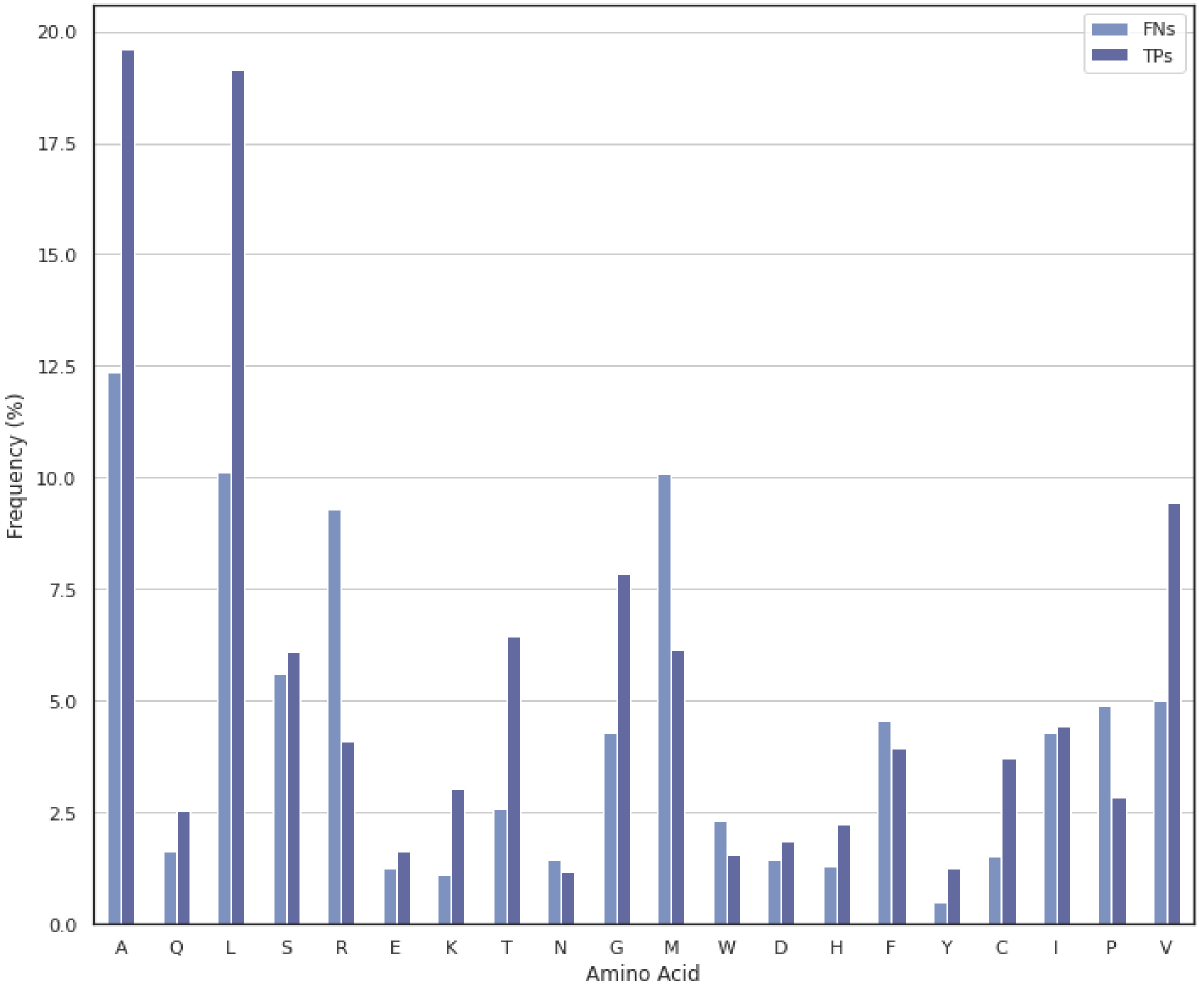

Fig. 6. Support vector machine false negative amino acid composition. The comparison of the false negative composition (FNs) is made with respect to the true signal peptide one (TPs). Particular evident is the decreased frequency of the most represented residues Leu (L) and Ala (A), the significant presence of hydrophobic residues Met (M), Trp (W), Phe (F), and the subsequential increment in the frequency of hydrophilic residues Arg (R), Pro (P), Asp (N).

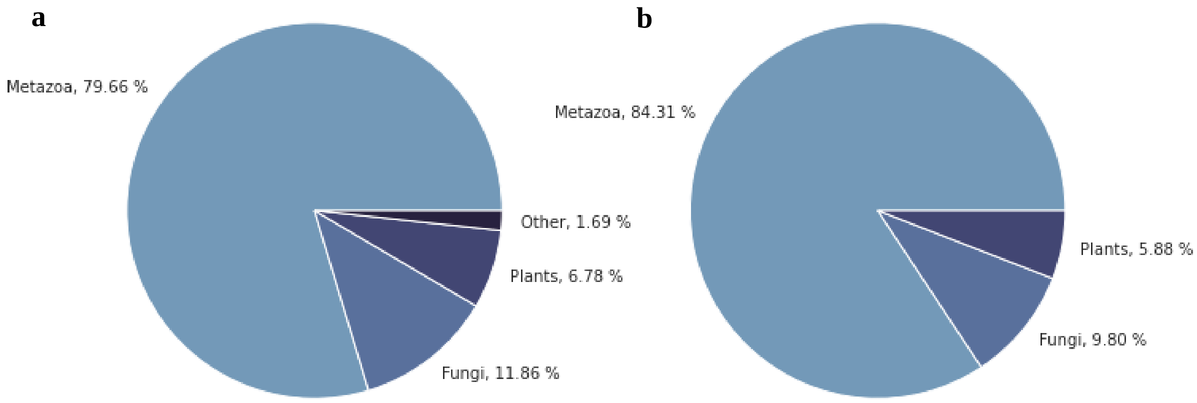

Fig. 7. False negative kingdom distribution. Panel **a** depicts the distribution in the von Heijne approach, panel **b** in the support vector machine. We can observe that the majority of false negative belong to the Metazoa kingdom.