Von Heijne and SVM (Support Vector Machine) models for the detection of signal peptides in UniProt protein sequences
Supplementary materials

Teresa Gianni, Martina Marotta, Bianca Mastroddi, Amedeo Antoci

Department of Pharmacy and Biotechnology - FaBiT, University of Bologna, Bologna, Italy

1 Data analysis

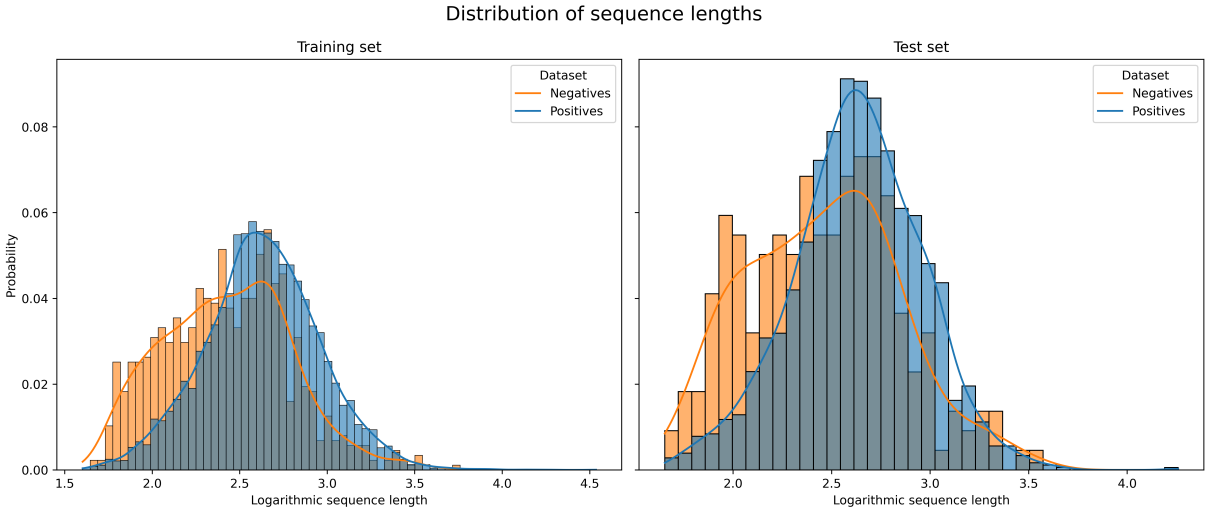


Figure 1: *The histograms show the distributions of sequence lengths between positive and negative entries in the training and test sets. The X-axis represents the logarithmic sequence length, and the Y-axis indicates the probability of observing sequences of a given length. In both sets, positive sequences tend to be longer than negative ones. A similar pattern was observed in the test set, indicating consistent behavior across datasets. Although the two distributions partially overlap, a shift toward longer sequence lengths in the positive class can be observed.*

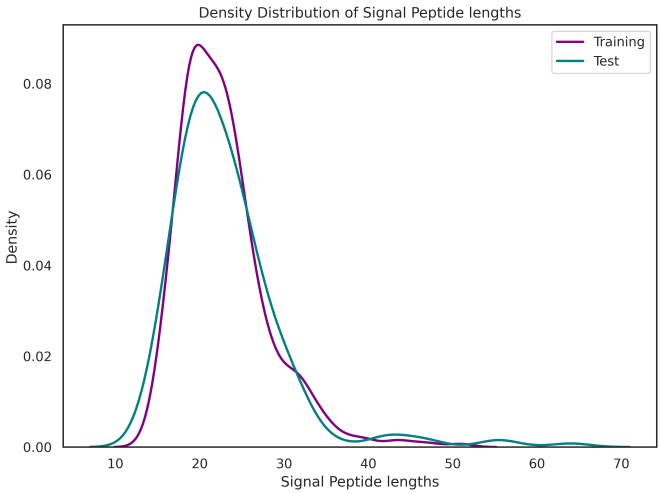


Figure 2: *The distribution of signal peptide lengths was assessed on the positive protein sequences, to verify the consistency between the training and test datasets. The comparison shows that the two density curves almost completely overlap, indicating a similar distribution of signal peptide (SP) lengths across both datasets. The training set exhibits a slightly higher density peak, suggesting a greater frequency of typical SP lengths. The peak of the distribution lies between approximately 15 and 30 amino acids, as expected from signal peptides.*

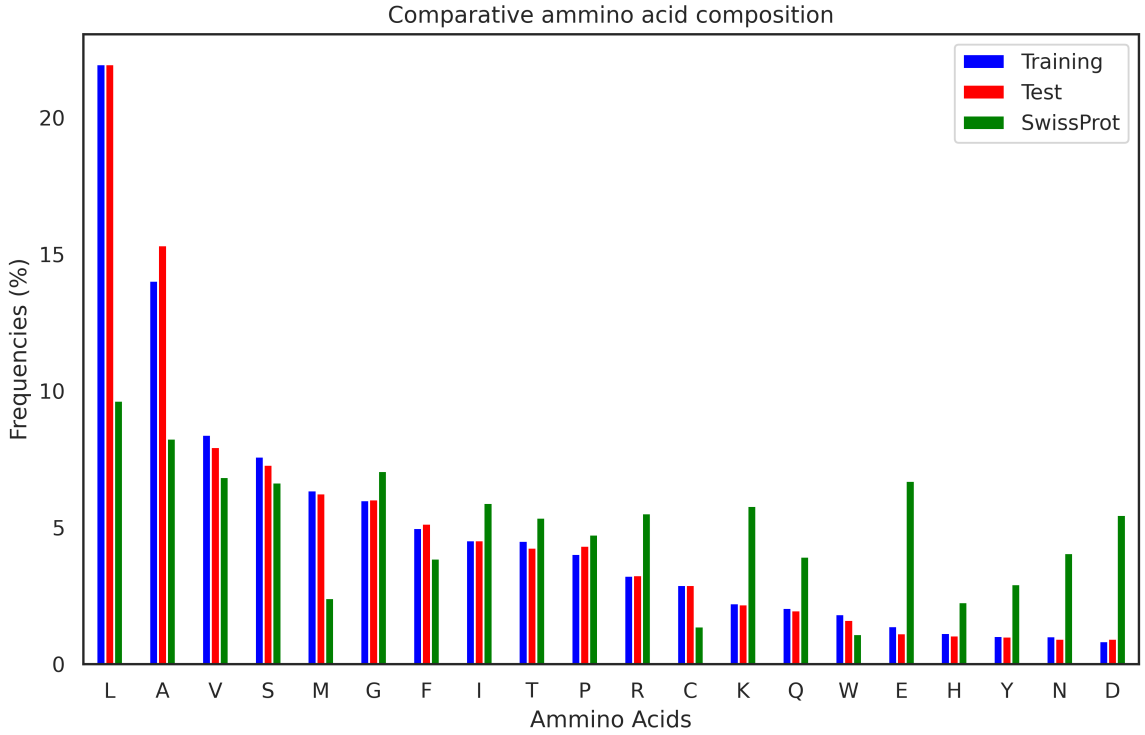


Figure 3: *The amino acid composition of signal peptides was compared among the training and test datasets and the general amino acid composition of all proteins in the SwissProt database, using only the positive protein sequences. The training and test sets displayed very similar amino acid distributions, consistent with the balanced nature of the datasets. Compared to the overall SwissProt distribution, signal peptides showed a higher frequency of hydrophobic residues,particularly leucine (L), which accounted for approximately 20% of residues compared to around 10% in the general SwissProt population. Other enriched residues included alanine (A), methionine (M), phenylalanine (F), and cysteine (C), which are characteristic of the hydrophobic core of signal peptides. Conversely, the relative frequency of acidic residues, such as glutamic acid (E) and aspartic acid (D), was markedly lower in the signal peptides.*

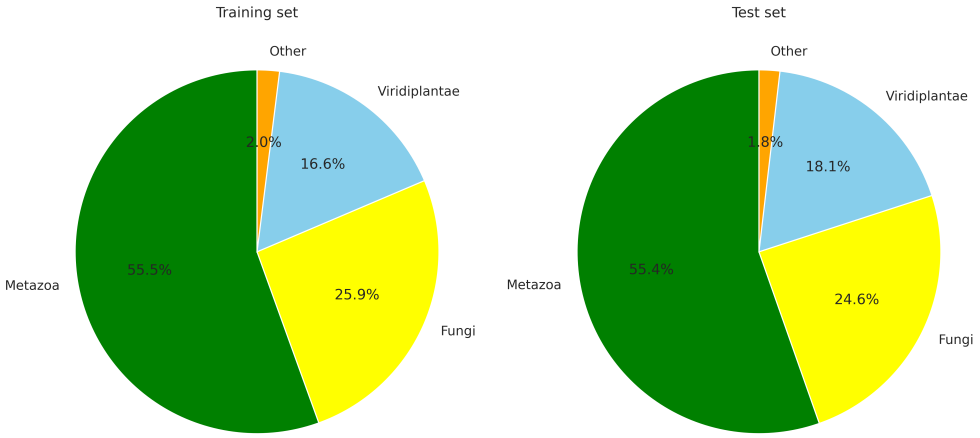
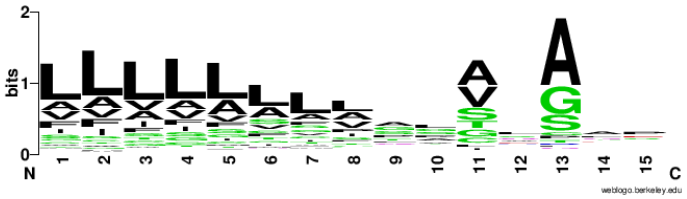


Figure 4: *The taxonomic distribution of protein sequences across kingdoms was evaluated for both the training and test datasets using pie charts. The majority of sequences belong to the Metazoa group, accounting for approximately 55% of the total, followed by Fungi, Viridiplantae, and a minor fraction classified as Other. Comparable proportions across both datasets indicate a balanced and representative split.*



(a) Training dataset.

(b) Test dataset

Figure 5: *Sequence logos were generated from the positive sequences of both the training and test datasets to analyze the frequency and conservation of residues around the signal peptide cleavage site (position -13 to +2). As expected, both datasets show an evident prevalence of leucine (L) and alanine (A) residues, consistent with the overall amino acid composition observed in the fig. 3*

2 Von Heijne model

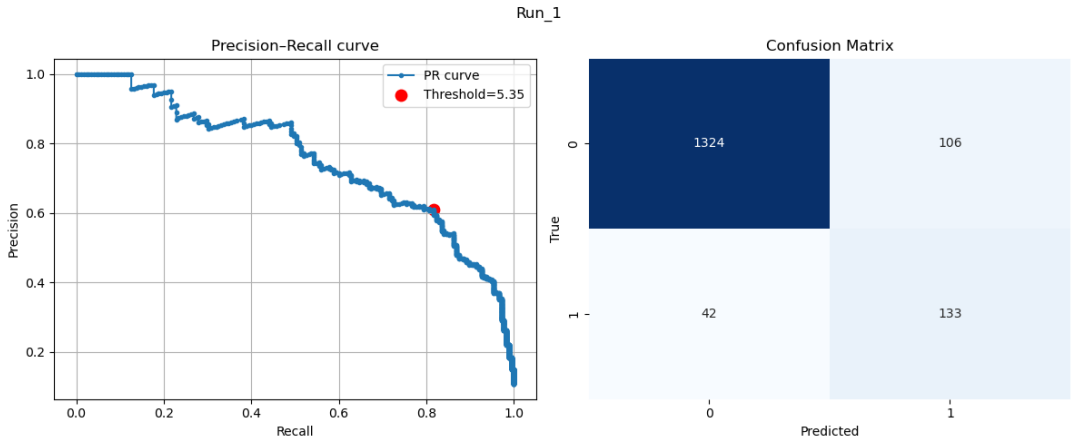


Figure 6: Run 1

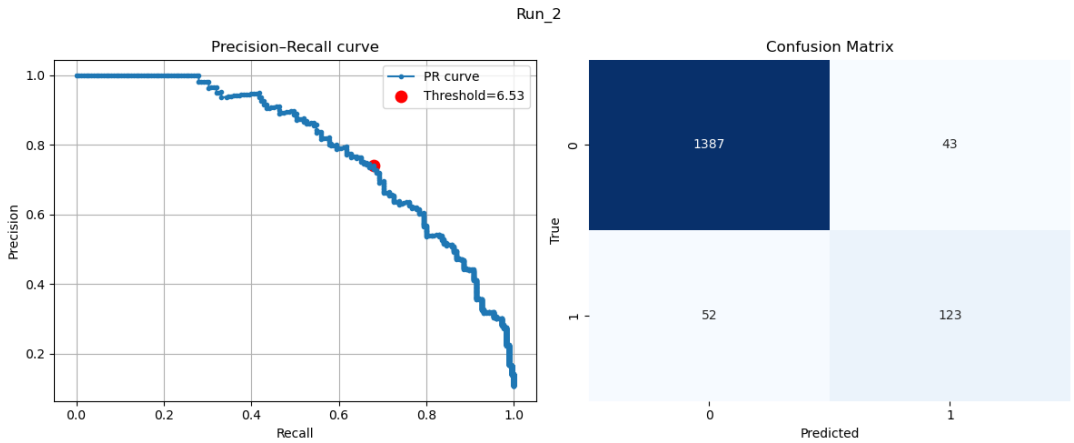


Figure 7: Run 2

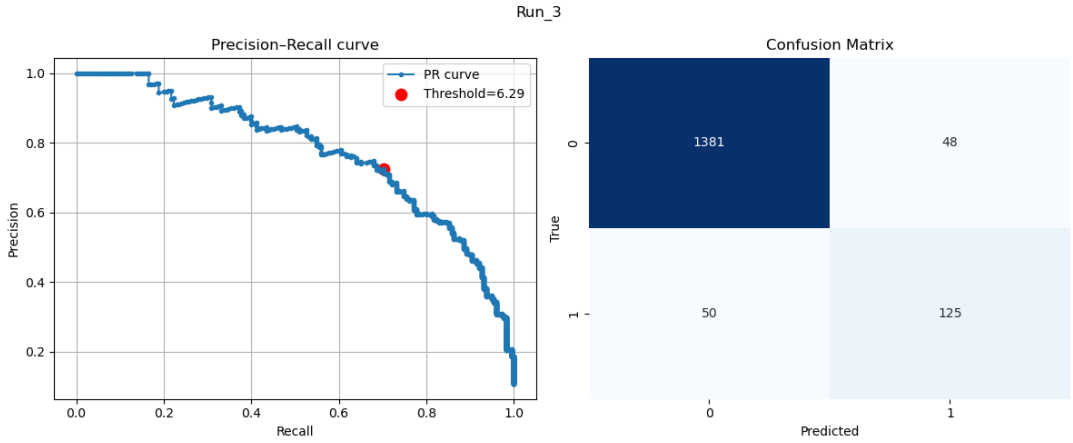


Figure 8: Run 3

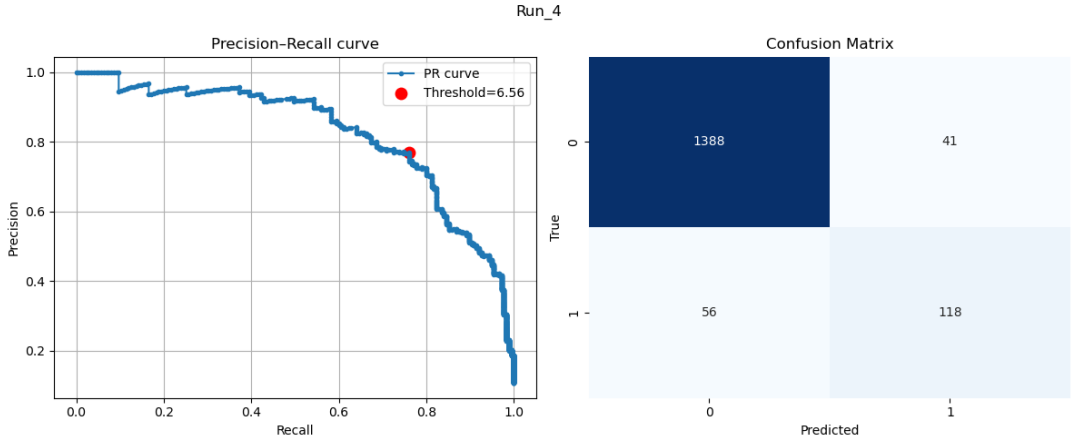


Figure 9: Run 4

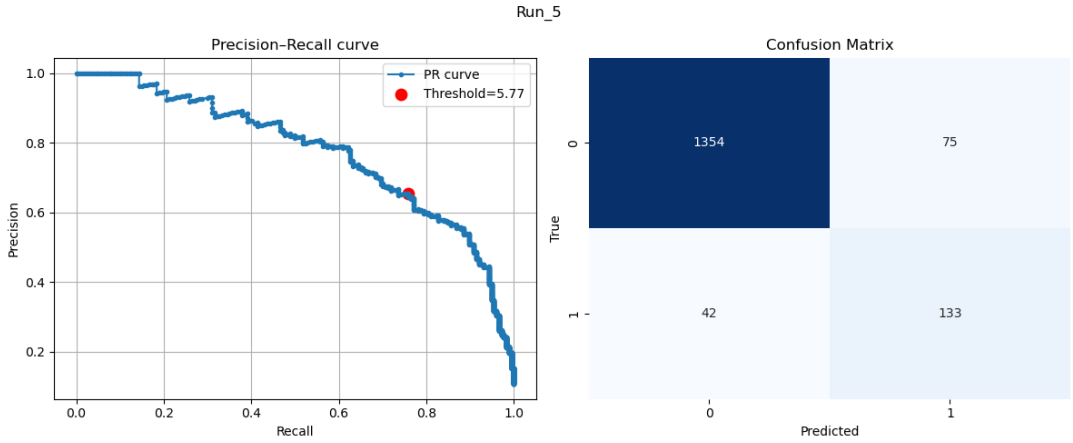


Figure 10: Run 5

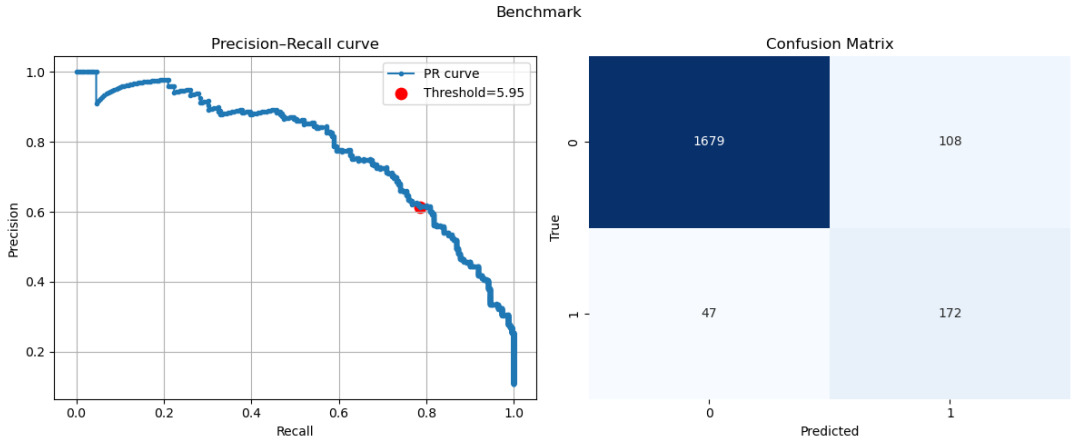


Figure 11: Benchmark

Figure 12: *Performance metrics of the Von Heijne model across five cross-validation runs (a-e) and the final evaluation on the Benchmark independent test set (f). Each panel displays the Precision-Recall (PR) Curve on the left and the corresponding Confusion Matrix on the right.*

3 SVM model

3.1 Gini importance

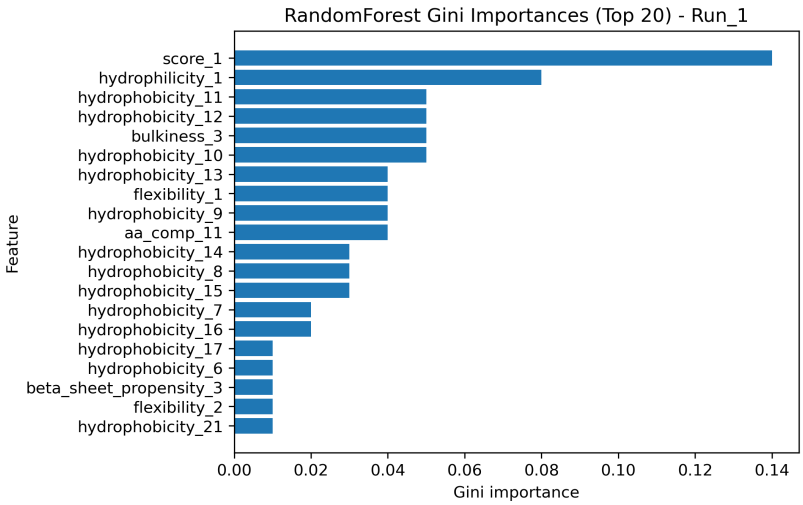


Figure 13: *Gini importances – Run 1*

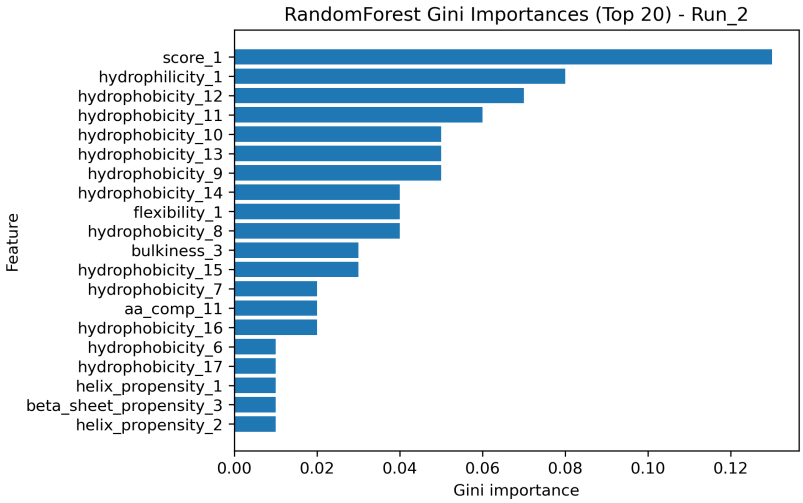


Figure 14: *Gini importances – Run 2*

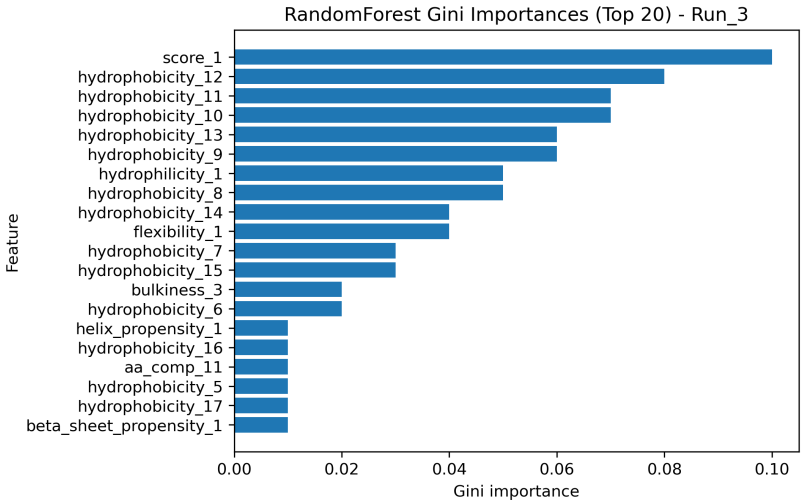


Figure 15: *Gini importances – Run 3*

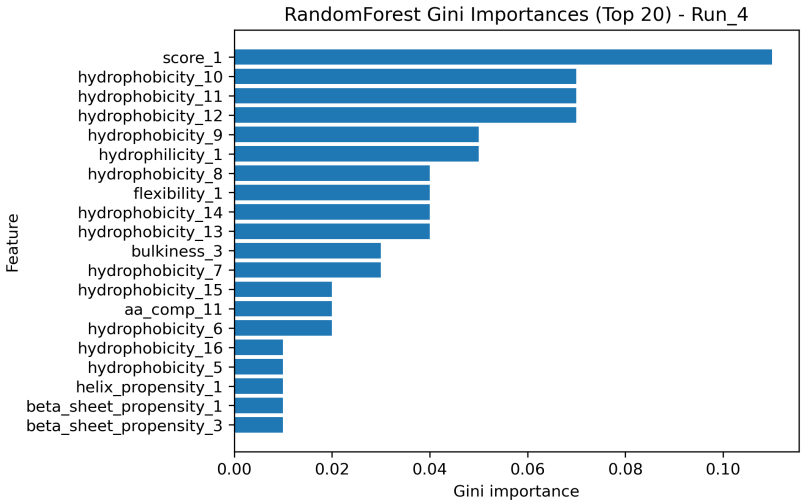


Figure 16: *Gini importances – Run 4*

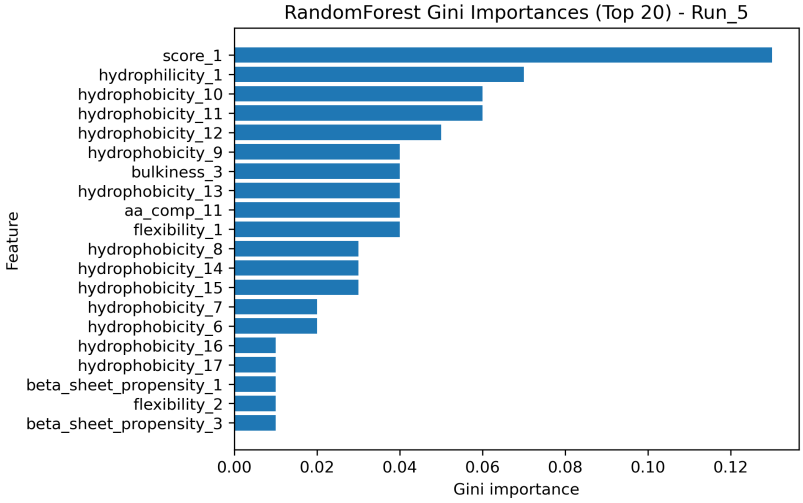
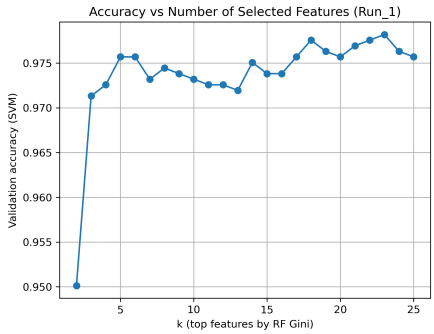
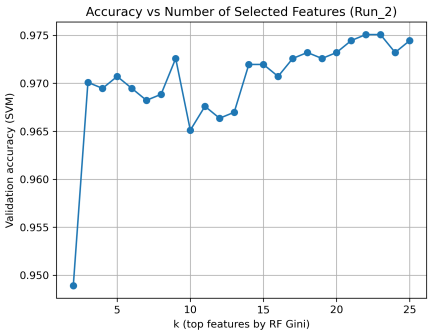


Figure 17: *Gini importances – Run 5*

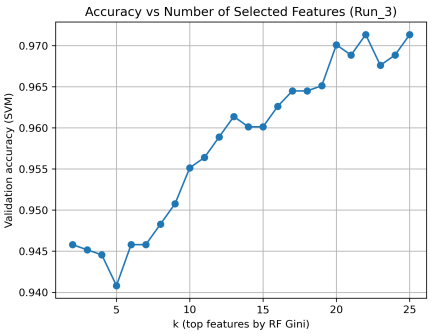
3.2 Accuracy per number of features



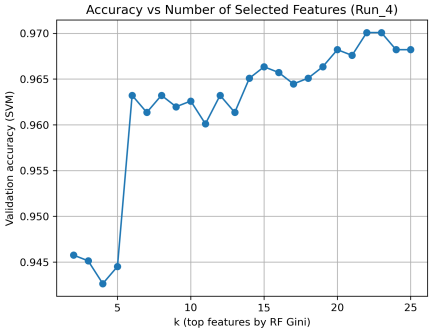
(a) *Accuracy vs features – Run 1*



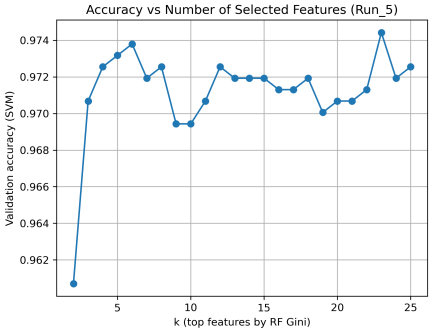
(b) *Accuracy vs features – Run 2*



(c) *Accuracy vs features – Run 3*



(d) *Accuracy vs features – Run 4*



(e) *Accuracy vs features – Run 5*

Figure 18: *Accuracy vs features among all the runs of cross-validation*

3.3 Confusion matrices

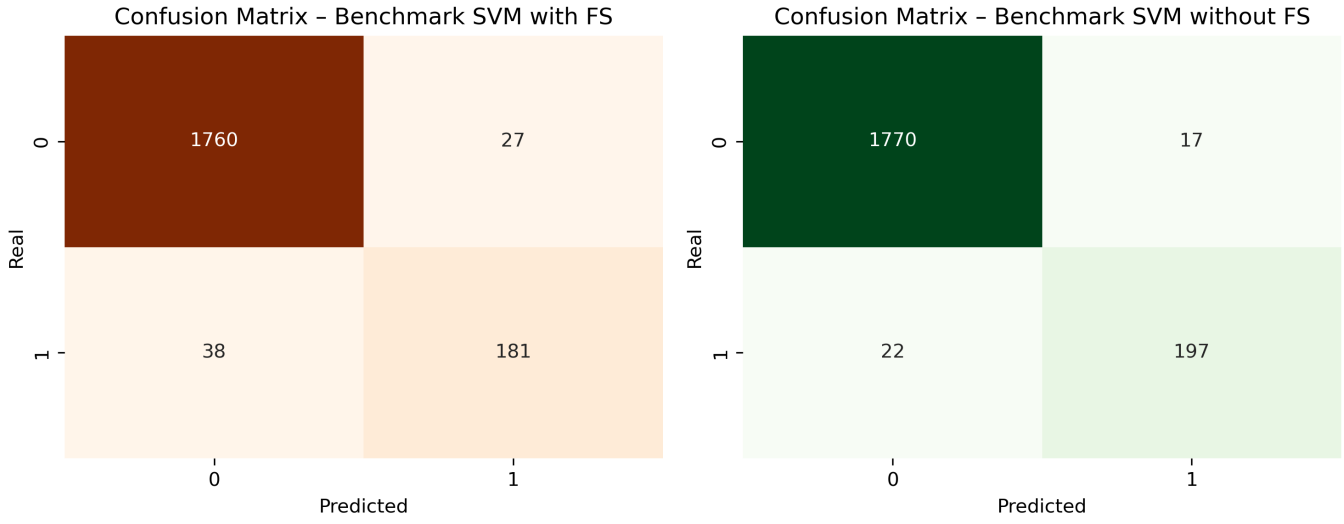
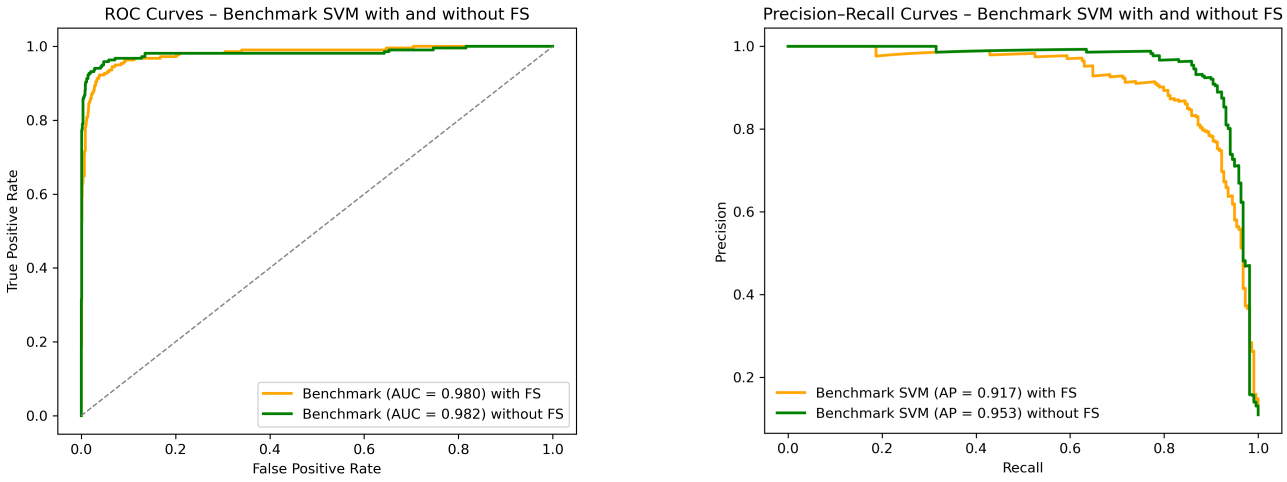


Figure 19: *Confusion matrices comparing the SVM model's predictive ability on the benchmark set with Feature Selection (FS) (left) and without FS (right). The matrix for the model without FS shows superior performance, yielding fewer False Positives (17 vs. 27) and fewer False Negatives (22 vs. 38).*

3.4 Benchmark performances



(a) **(b)**

Figure 20: ***(a)*** *Comparison of Receiver Operating Characteristic (ROC) curves between the SVM model with Feature Selection (orange) and the final model without Feature Selection (green) on the benchmark test set. Both models exhibit excellent discriminative power with near-perfect Area Under the Curve (AUC) values: AUC=0.980 for the model with FS and AUC=0.982 for the model without FS. The similarity of the curves suggests robust overall separability of the classes by the SVM classifier.*
(b) *Comparison between Precision-Recall curves between SVM model with (orange) and without (green) Feature-Selection*

4 False positives false negatives analysis

4.1 SVM features

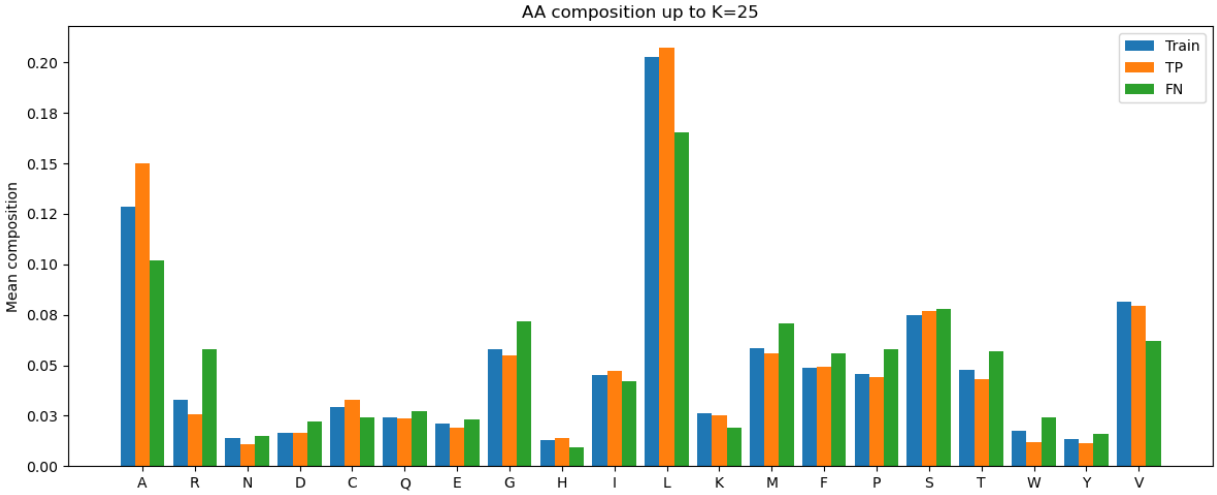


Figure 21: *Comparative amino acid composition (K=25) of True Positives (TP), False Negatives (FN) and train examples from the SVM model. The plot shows that FN sequences (misclassified SPs) often have a lower mean composition of key hydrophobic residues like Leucine (L) and Alanine (A) compared to the correctly classified TP sequences, suggesting that FN sequences are likely atypical SPs with reduced hydrophobicity.*

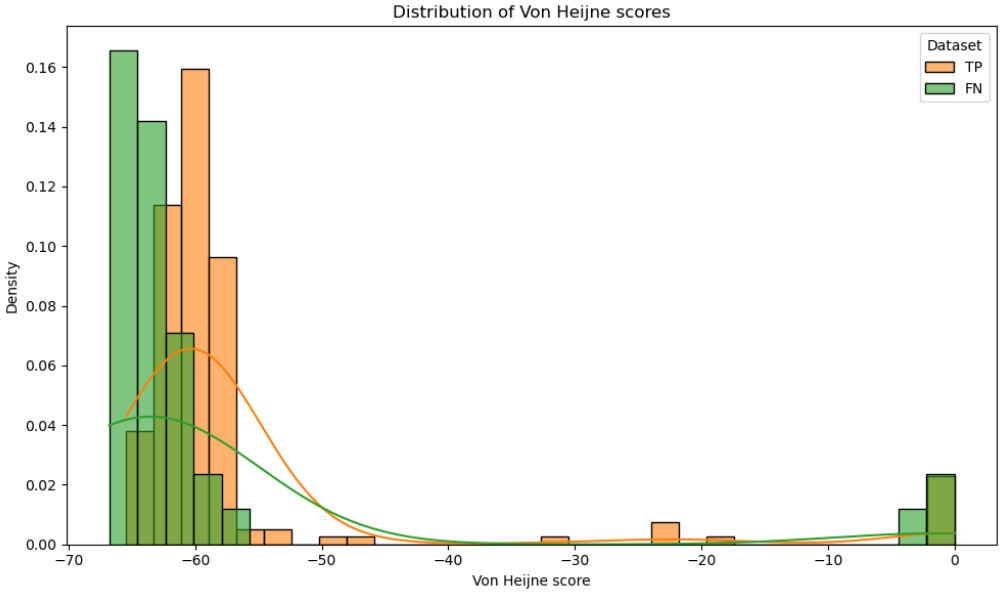


Figure 22: *Density distribution of Von Heijne scores for True Positives (TP) and False Negatives (FN) from the benchmark set. While TP sequences generally cluster around higher scores (closer to 0), a significant fraction of FN sequences (true SPs missed by the model) show lower, more diffuse scores, reinforcing the idea that the Von Heijne model struggles with atypical signal peptides that deviate from the standard PSWM motif.*

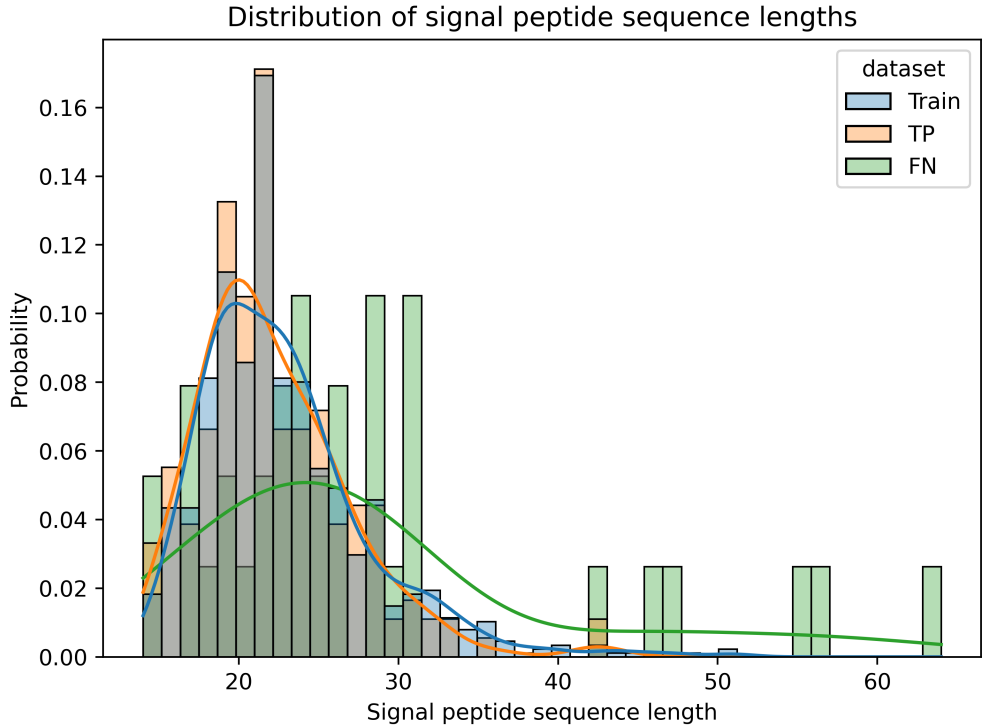


Figure 23: *Comparative distribution of signal peptide lengths for the Training (Train) set, True Positives (TP), and False Negatives (FN). The plot highlights that FN sequences have a broader and slightly shifted distribution, including a higher proportion of longer SPs compared to the typical SP lengths (15-30 aa) found in the Train and TP sets.*

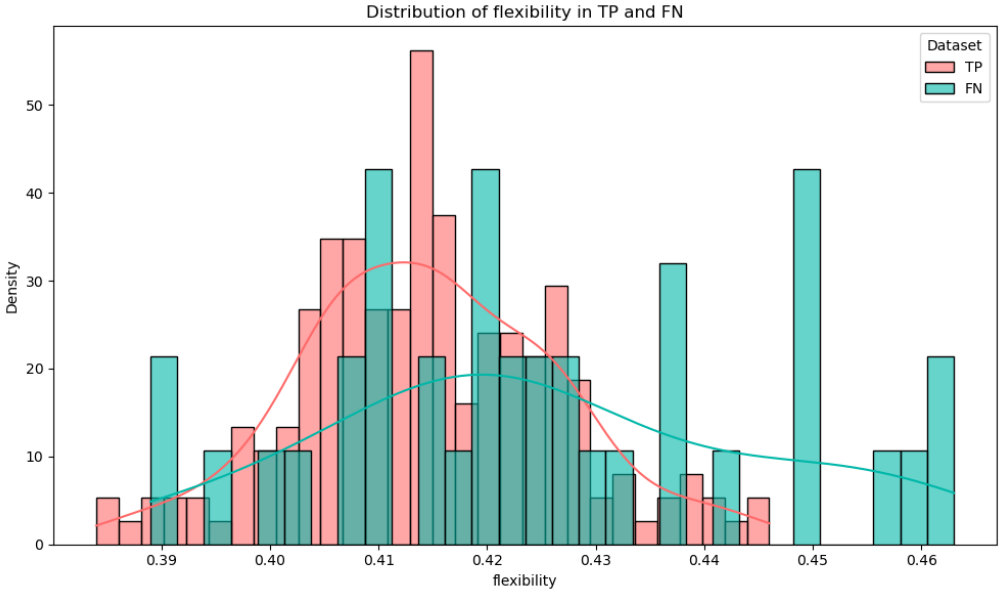


Figure 24: *Distribution of the Flexibility feature scores for True Positives (TP) and False Negatives (FN). The chart shows that while TP sequences are generally concentrated around the mean flexibility value, the distribution for FN sequences is broader, suggesting that misclassified signal peptides may possess higher or more variable flexibility, contributing to the model's difficulty in correctly identifying them.*

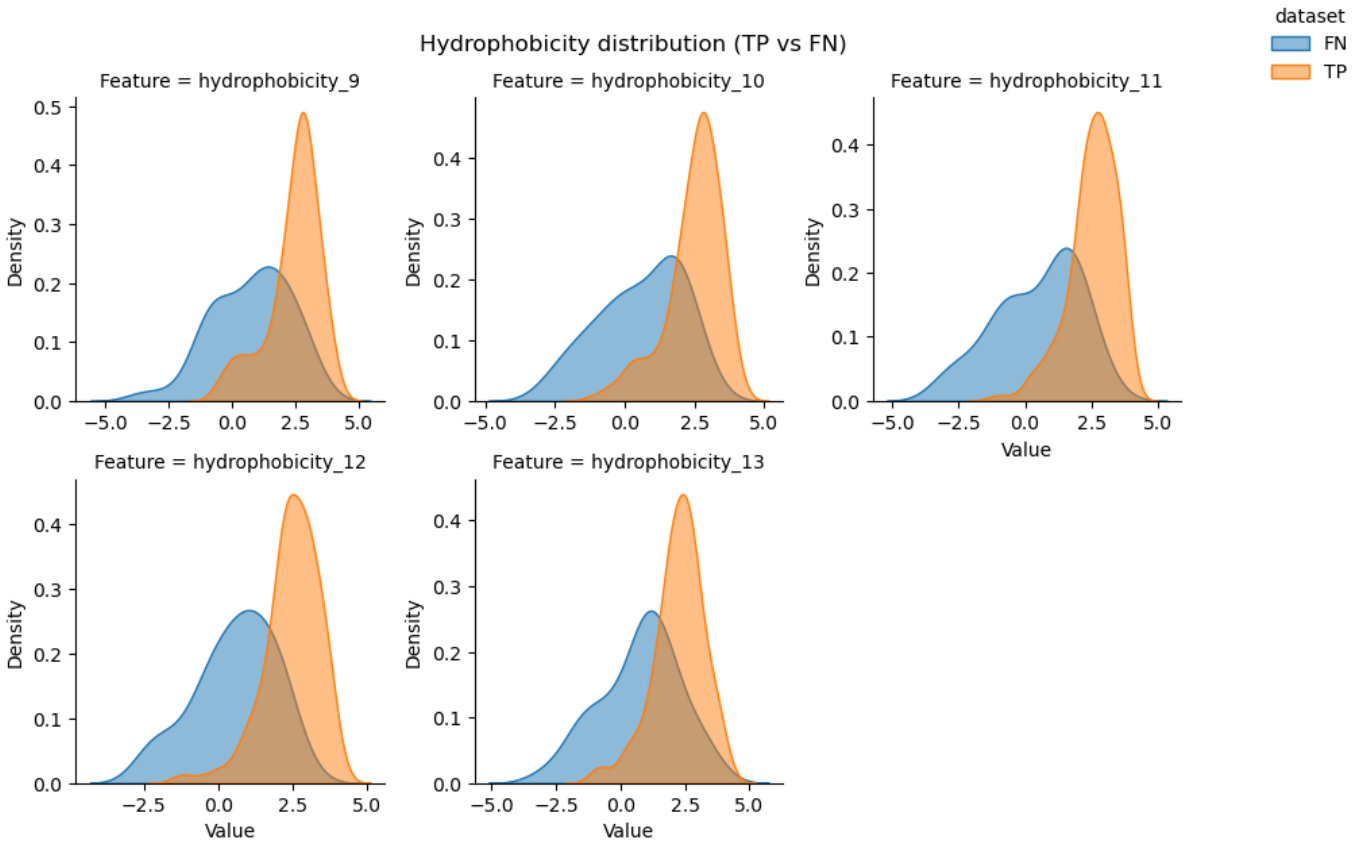


Figure 25: *Density distributions of the Hydrophobicity feature across five crucial positions (positions 9 to 13 relative to the sequence start) for the SVM's classes. The clear separation and non-overlapping peaks between the classes (representing SP vs. non-SP) at these positions demonstrate the high discriminative power of the hydrophobicity features in the SVM model.*