Supplementary Materials LB2 Group 5

Learning the Secretory Code: A Comparative Study of Von Heijne, SVM, and MLP Models for Signal Peptide Classification

Alessia Corica, Anna Rossi, Martina Castellucci, Sofia Natale

Overview

This document serves as the essential technical companion to the main study, "Learning the Secretory Code: A Comparative Study of Von Heijne, SVM, and MLP Models for Signal Peptide Classification". Contained within are detailed figures and tables that illustrate the computational pipeline supporting the methods and results presented in the report. Additional explanations and procedural clarifications are provided in the Supplementary Materials.

1 Supplementary Data Analysis

This section reports additional exploratory analyses performed on the curated training and benchmark datasets. The aim is to verify that the sequence collections are biologically coherent and statistically comparable, excluding major biases in protein length, signal peptide length, amino acid composition, and taxonomic representation, and to visualise the canonical cleavage-site motif of eukaryotic signal peptides.

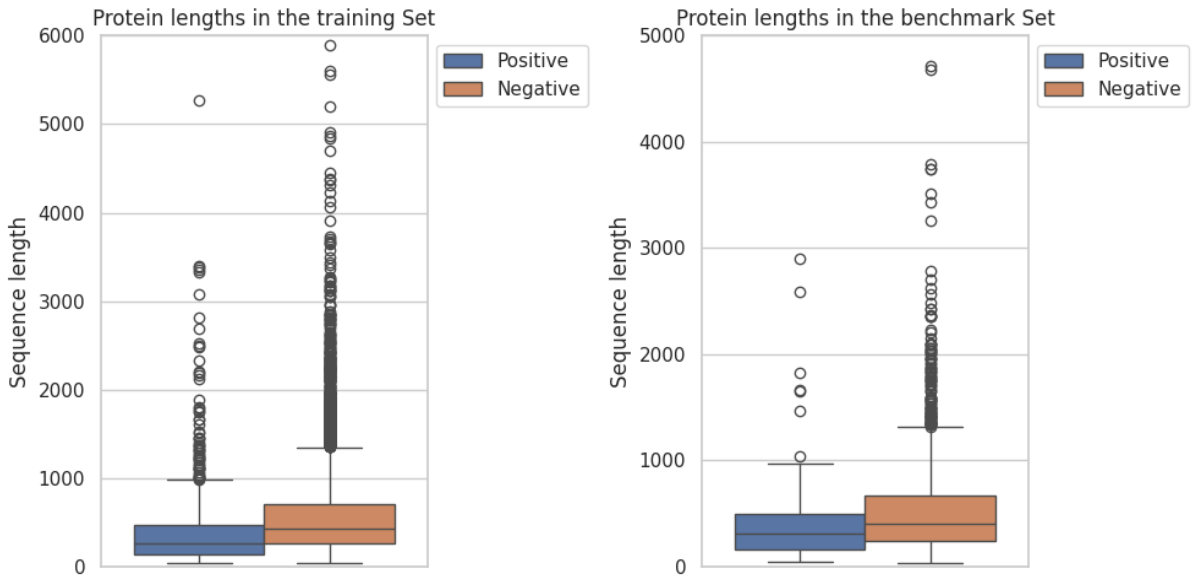

Figure 1: Protein length distributions in the training (left) and benchmark (right) sets. Boxplots compare positive (SP-containing) and negative proteins. Both datasets show broadly similar ranges and medians, indicating that the models are not trivially driven by global sequence length differences.

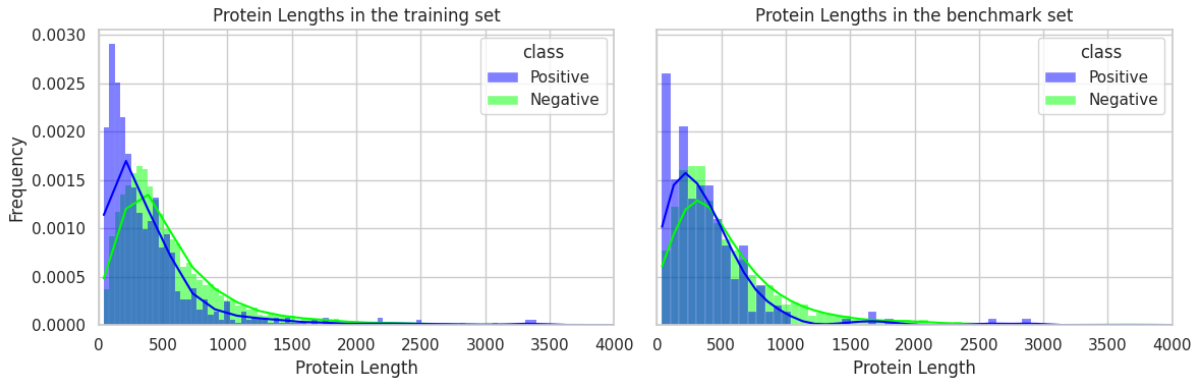

Figure 2: Kernel density estimates of protein length in the training (left) and benchmark (right) sets. Positive and negative classes display overlapping distributions with a shared right tail, confirming that the curated datasets retain realistic variability in sequence size.

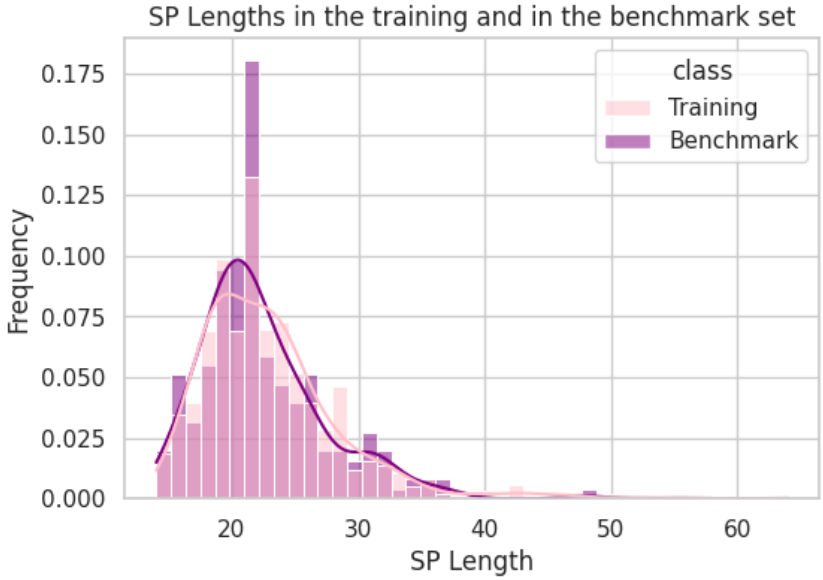

Figure 3: Signal peptide length distribution in the training and benchmark sets. Most SPs fall in the 18–30 amino acid range, with a long but sparse tail of longer sequences. The close overlap between the two curves indicates that the benchmark set is representative of the training distribution.

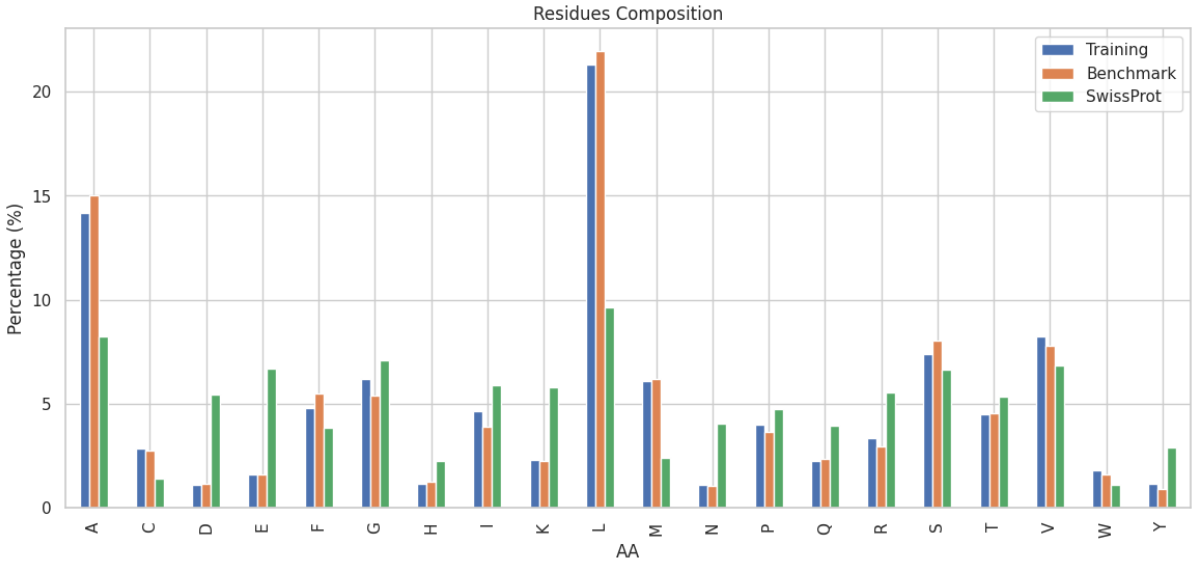

Figure 4: Amino acid composition of signal peptides in the training and benchmark sets compared with SwissProt background frequencies. Both datasets show the expected enrichment in hydrophobic residues (A, L, V, I, M) and depletion of charged and aromatic residues, confirming that the curated SPs preserve the canonical compositional bias of eukaryotic signal peptides.

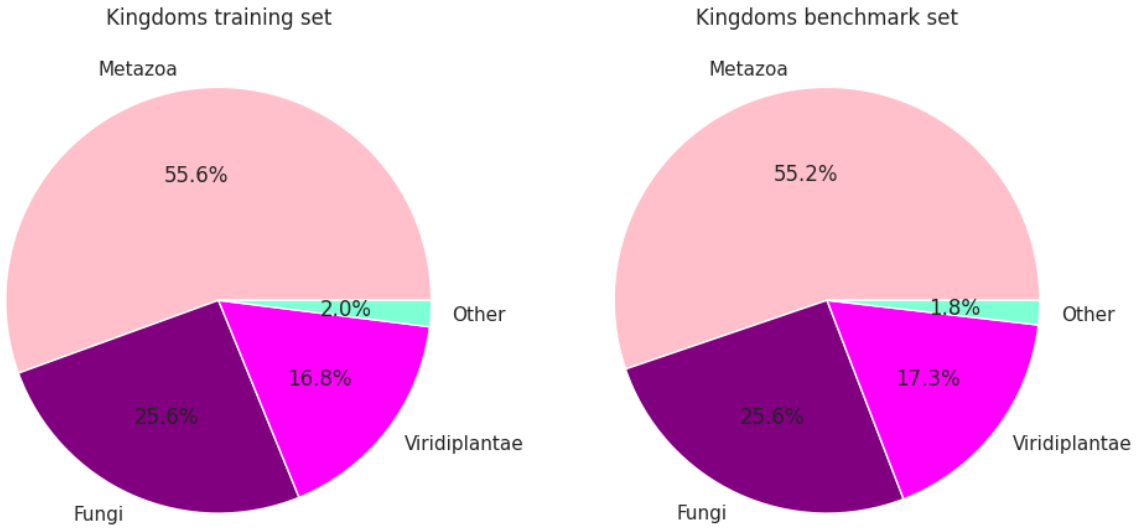

Figure 5: Kingdom-level distribution of source organisms in the training (left) and benchmark (right) sets. Metazoa, Fungi, and Viridiplantae dominate both collections with very similar proportions, indicating that taxonomic diversity is preserved between training and evaluation datasets.

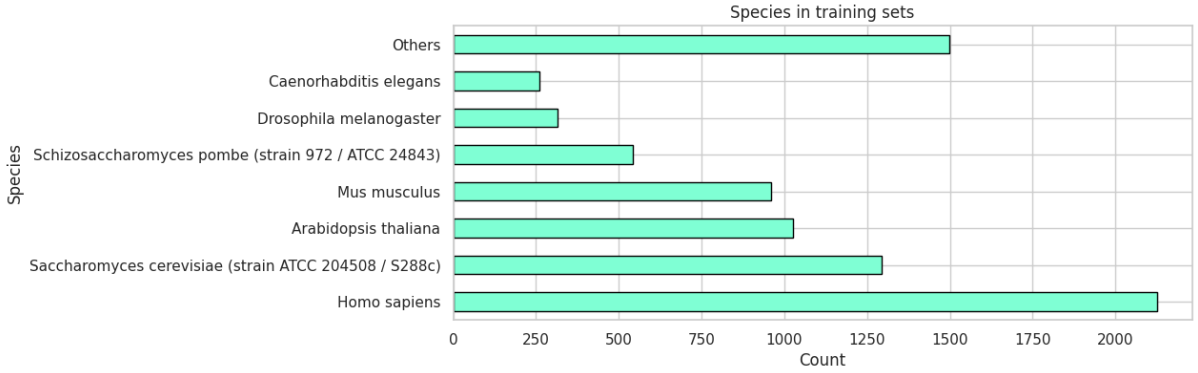

Figure 6: Major contributing species in the positive training set. Homo sapiens, Saccharomyces cerevisiae, Arabidopsis thaliana, and Mus musculus account for most entries, while a heterogeneous "Others" category aggregates less represented taxa. This distribution reflects the annotation density of model organisms in UniProtKB.

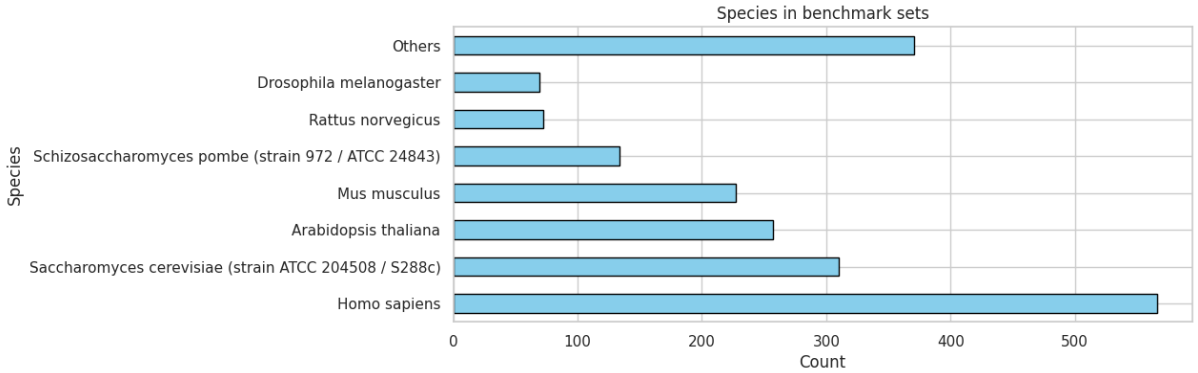

Figure 7: Major contributing species in the positive benchmark set. Homo sapiens and Saccharomyces cerevisiae remain the most represented taxa, followed by Arabidopsis thaliana and Mus musculus, while a heterogeneous "Others" category aggregates less frequent species. The overall distribution closely mirrors that of the training set, indicating that benchmark evaluation is performed on a taxonomically comparable subset.

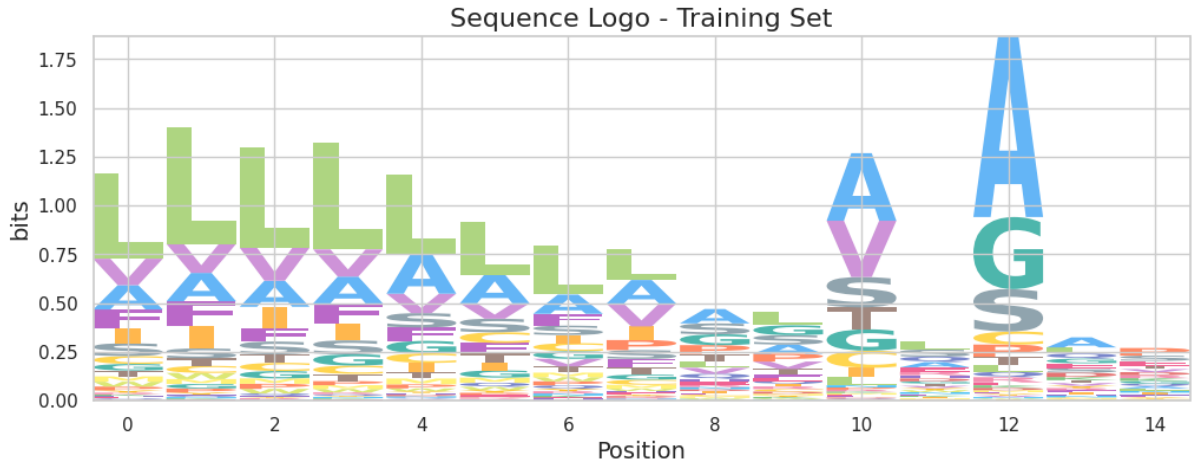

Figure 8: Sequence logo for the $[-13, +2]$ window around the signal peptidase cleavage site in the training set. The logo shows a hydrophobic H-region upstream and strong enrichment of small residues at positions -1 and -3, consistent with the canonical signal peptide motif.

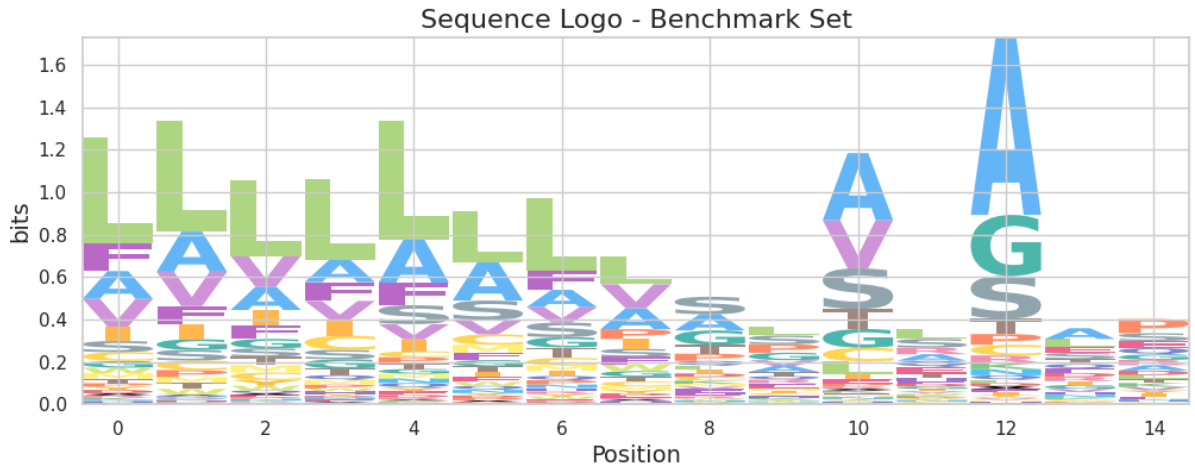

Figure 9: Sequence logo for the $[-13, +2]$ window in the benchmark set. The pattern closely matches the training logo, confirming that motif strength and residue preferences are comparable between training and evaluation data.

2 Supplementary Von Heijne Method

This section reports the supplementary analyses supporting the von Heijne statistical classifier, including the PSWM characterization, threshold optimization across the five cross-validation folds, fold-wise confusion matrices, the aggregated PR-curve comparison, and distribution of benchmark species.

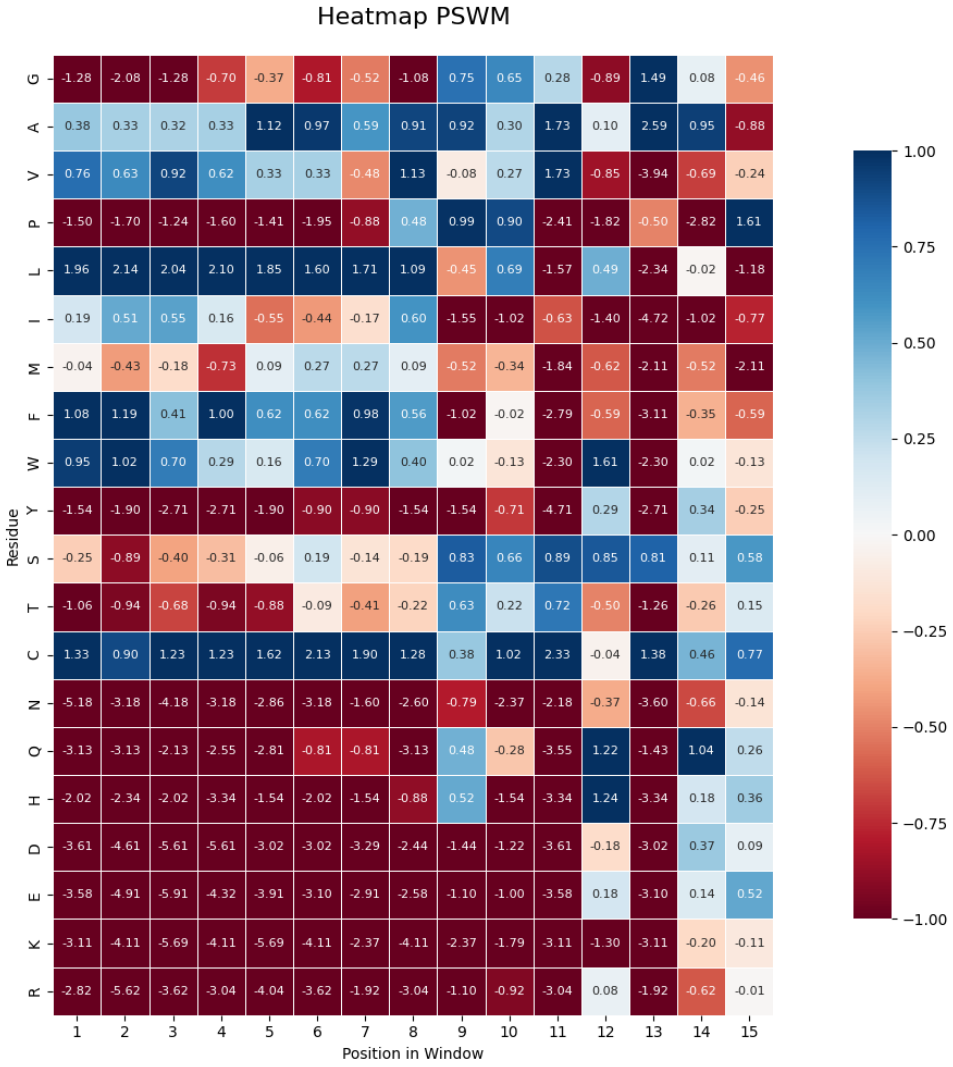

Figure 10: PSWM log-odds heatmap used by the von Heijne model. Each cell represents the log-odds enrichment of an amino acid at a specific position within the 16-residue window (13 to +2). Positive (blue) values indicate residues more frequent than expected from SwissProt background frequencies, while negative (red) values indicate depletion. A strong hydrophobic core is visible between positions $1 - 8$, while the characteristic `A/G` pattern at positions $+1/+2$ reflects the canonical signal peptidase cleavage motif.

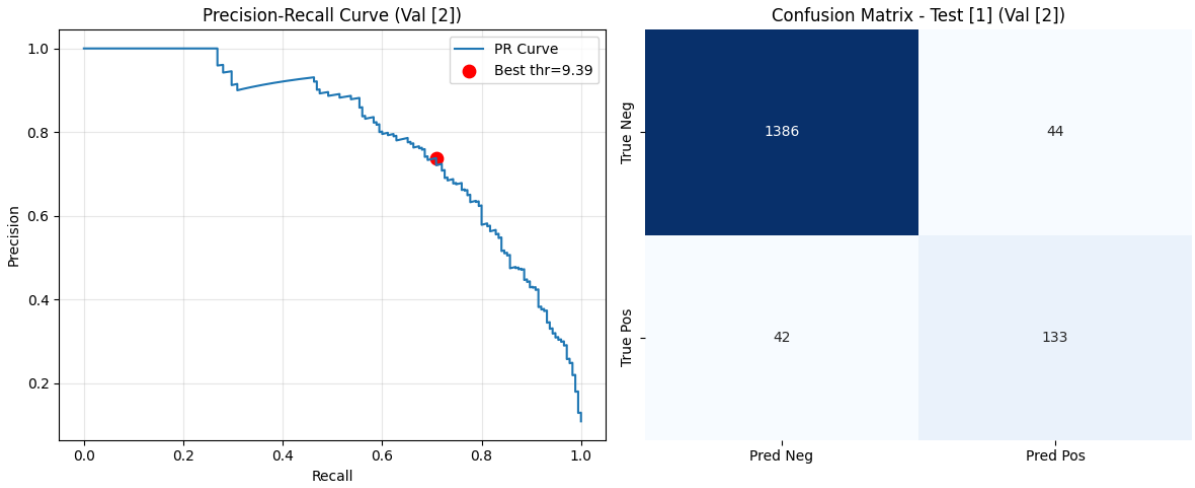

Figure 11: Fold 1 — Precision–Recall curve and corresponding confusion matrix. The red marker identifies the threshold (thr = 9.39) maximizing the F1-score on the validation split. The confusion matrix shows a balanced distribution of false positives and false negatives, reflecting the intrinsic sensitivity of PSWM scoring to hydrophobicity variations.

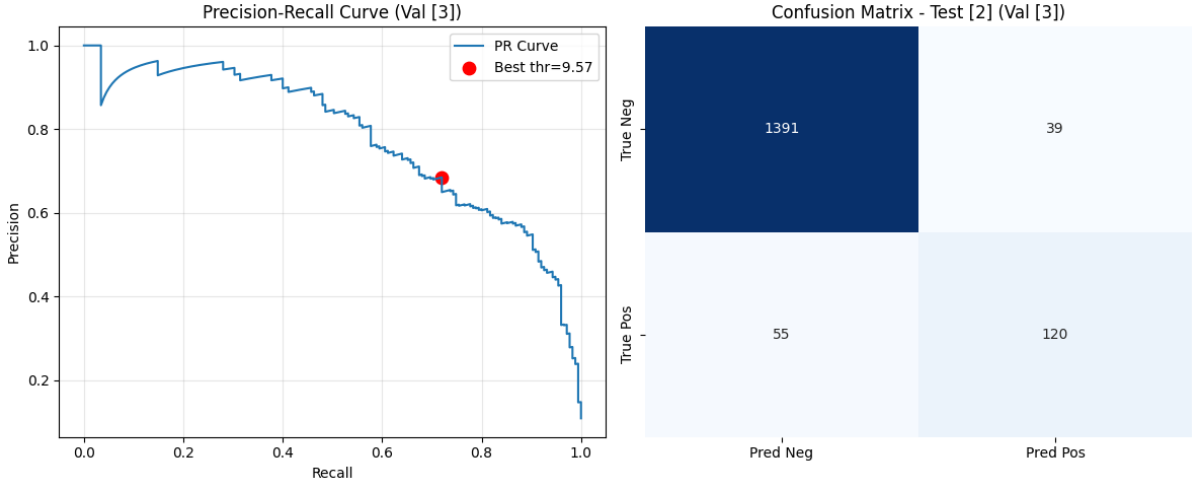

Figure 12: Fold 2 — PR curve and confusion matrix. The optimal threshold (thr = 9.57) maintains high precision across most recall values. Misclassifications in this fold mainly arise from sequences with weakened H-region hydrophobicity.

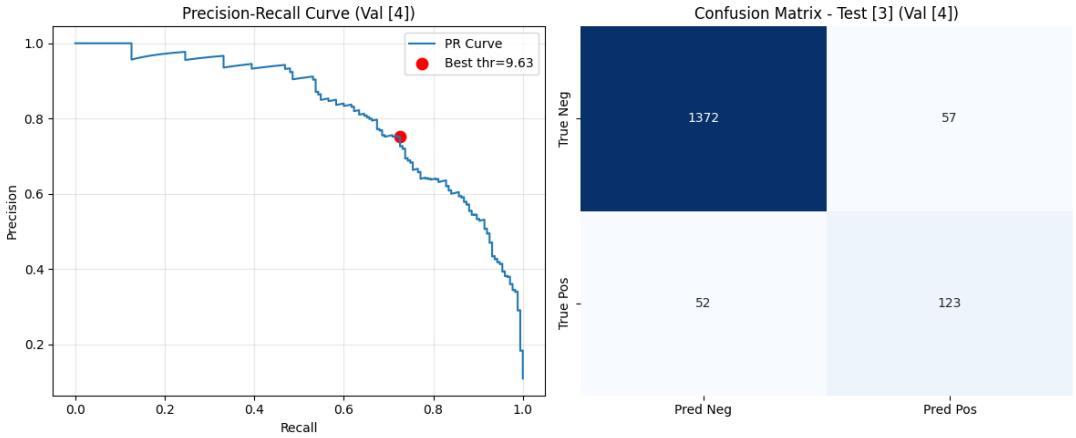

Figure 13: Fold 3 — PR curve and confusion matrix. At the selected threshold (thr = 9.63), the model favors recall over precision. False positives predominantly correspond to N-terminal regions resembling hydrophobic transmembrane segments.

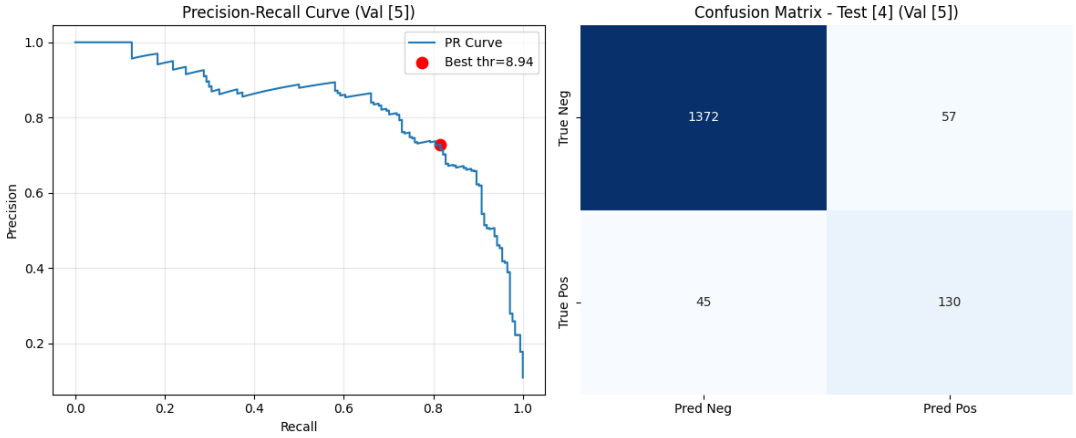

Figure 14: Fold 4 — PR curve and confusion matrix. The threshold (thr = 8.94) is the lowest across folds, indicating that this subset contains SPs with weaker overall PSWM scores. The confusion matrix shows increased FN counts, driven by shorter SPs (<18 aa).

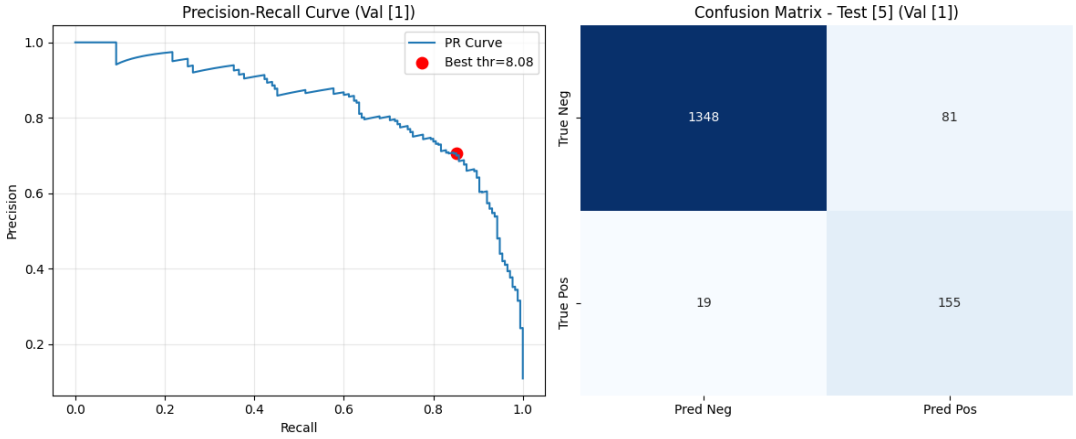

Figure 15: Fold 5 — PR curve and confusion matrix. The optimal threshold (thr = 8.08) is associated with the highest recall across folds. False positives reflect high-scoring but non-secretory hydrophobic segments.

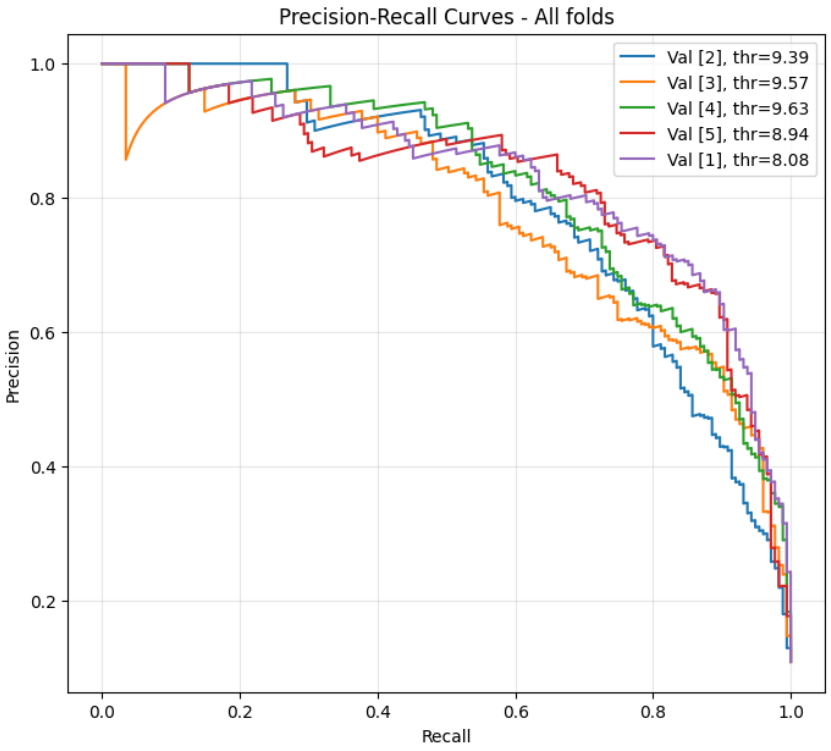

Figure 16: Superimposed Precision–Recall curves for all cross-validation folds. Curves show consistent performance across folds, with recall >0.7 at precision 0.7 for most splits. Variability at low recall values originates from fold-specific motif composition and hydrophobic core length differences. The curve cluster demonstrates the stability of the von Heijne scoring scheme despite being motif-dependent and non-parametric.

3 Supplementary SVM Method

This section reports additional figures supporting the feature–engineering, model selection, and evaluation process for the Support Vector Machine (SVM) classifier. Each plot highlights a specific aspect of the training workflow, from feature importance to performance diagnostics.

Table 1: Optimal SVM configuration and evaluation results.

Hyperparameters	**Value**
C	10
Kernel	RBF
Gamma	scale
Results	**Score**
Best CV F1	0.720
Validation Accuracy	0.922
Selected Features	15

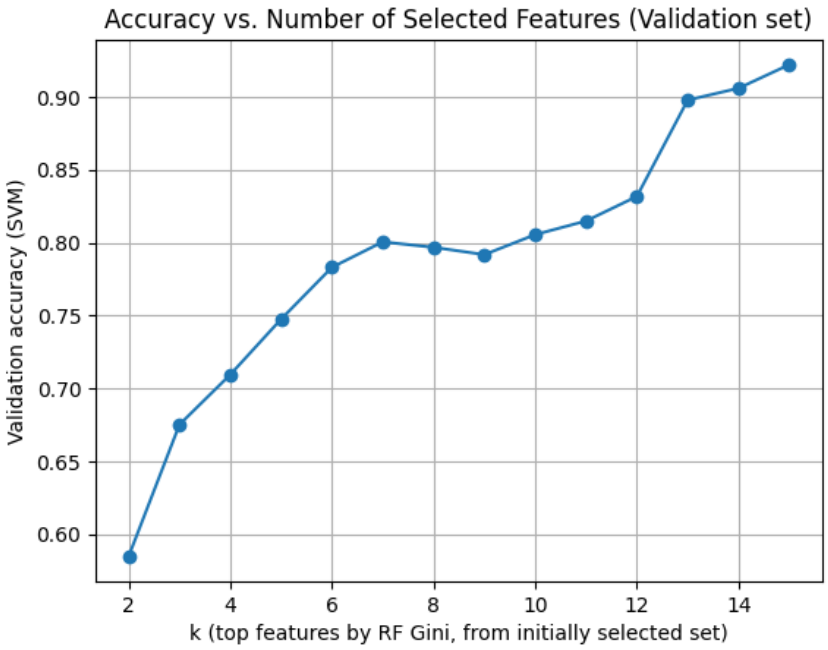

Figure 17: Validation accuracy as a function of the number of top features selected by Random Forest. The curve shows progressive improvement up to $k = 15$ features, where accuracy reaches its maximum (0.922). This supports the choice of retaining 15 physicochemical descriptors for downstream SVM training.

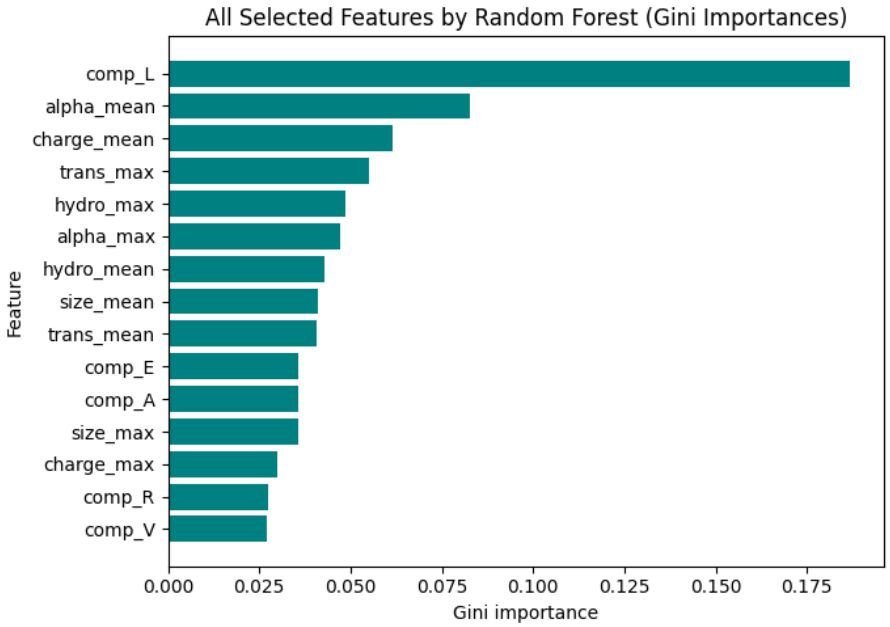

Figure 18: Random Forest Gini importance ranking for all extracted features. Features related to hydrophobicity, charge, secondary–structure propensity and sequence composition dominate the importance spectrum. Notably, `comp_L` is the strongest individual predictor, confirming the biological relevance of Leucine enrichment in signal peptides.

Table 2: Gini-based feature importance ranking computed from the Random Forest used for feature selection. From 29 initial features, 15 were retained above the median importance threshold.

Rank	**Feature**	**Importance**
1	comp_L	0.187
2	α-helix mean	0.083
3	Charge mean	0.062
4	Transmembrane max	0.055
5	Hydrophobicity max	0.049
6	α-helix max	0.048
7	Hydrophobicity mean	0.043
8	Size mean	0.041
9	Transmembrane mean	0.041
10	comp_E	0.036

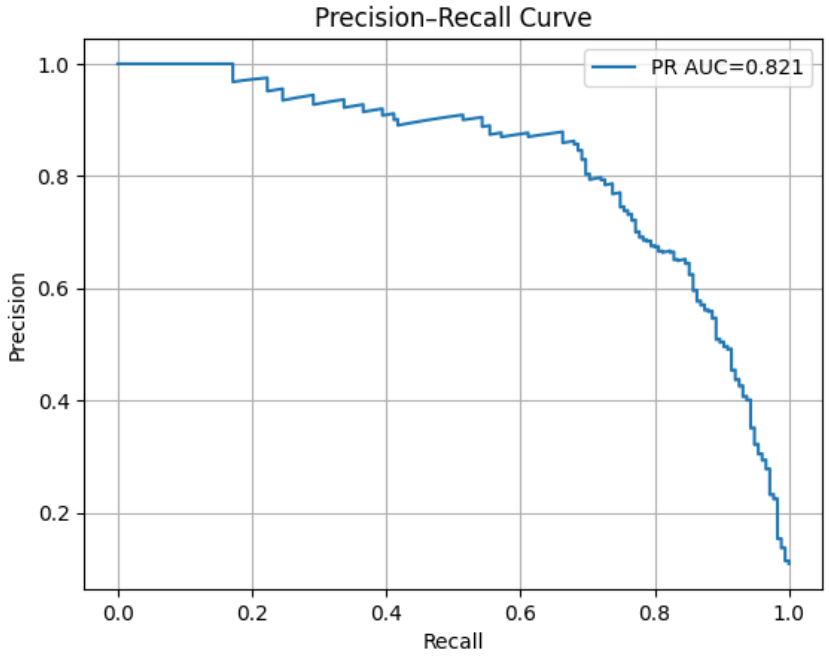

Figure 19: Precision–Recall curve for the best SVM model. The area under the curve (PR–AUC = 0.821) indicates robust discrimination between SP and non-SP sequences even under class imbalance. Precision remains above 0.85 for recall values up to 0.6, demonstrating stable reliability across operating thresholds.

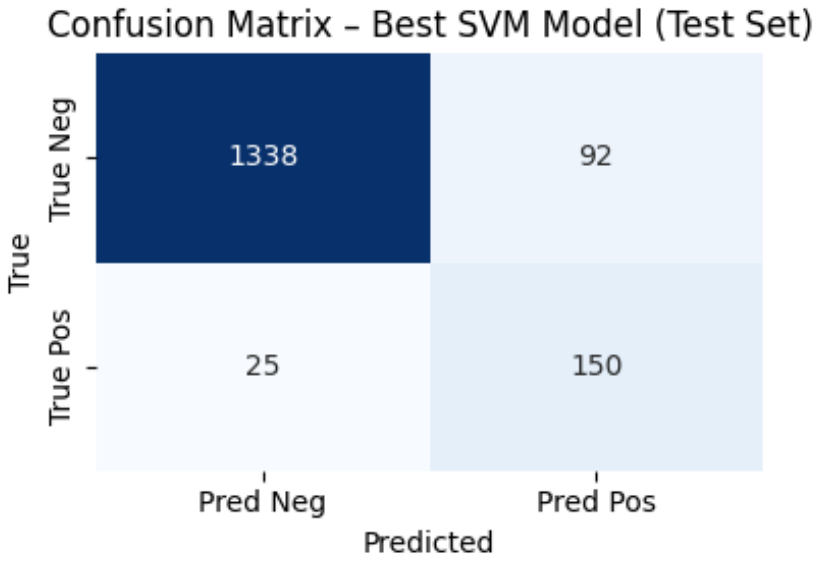

Figure 20: Confusion matrix for the best SVM model on the internal validation/test split. The classifier correctly identifies most negative samples (1338 TN) and retrieves a large fraction of positives (150 TP), with relatively few false classifications (92 FP, 25 FN). This balance supports good precision–recall trade-off.

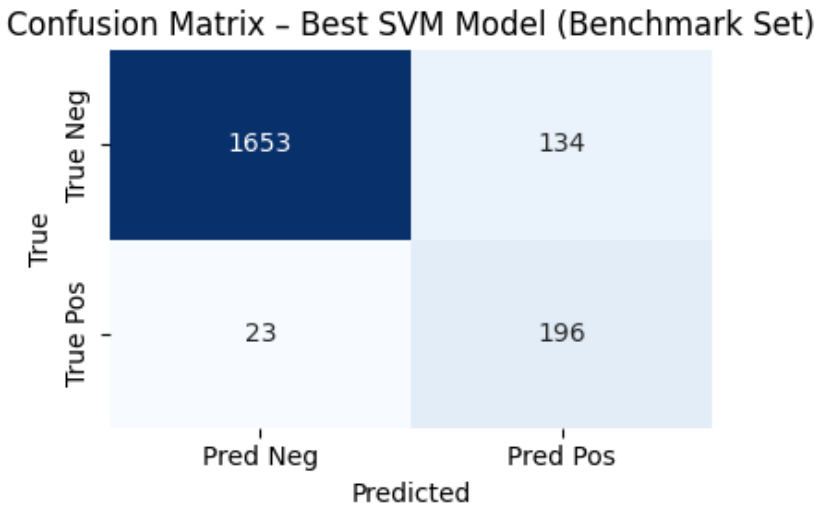

Figure 21: Confusion matrix for the independent benchmark dataset. Performance generalizes well to unseen sequences, with 1653 TN and 196 TP. The number of false positives (134) is consistent with the hydrophobicity-driven tendency to misclassify transmembrane N-terminal regions.

4 Supplementary Evaluation and Error Analysis

This section reports additional analyses supporting the comparative evaluation of the von Heijne and SVM models. It summarises the main error patterns, highlights differences between true and misclassified signal peptides, and examines key factors such as hydrophobicity, charge, transmembrane propensity, motif conservation, and signal peptide length.

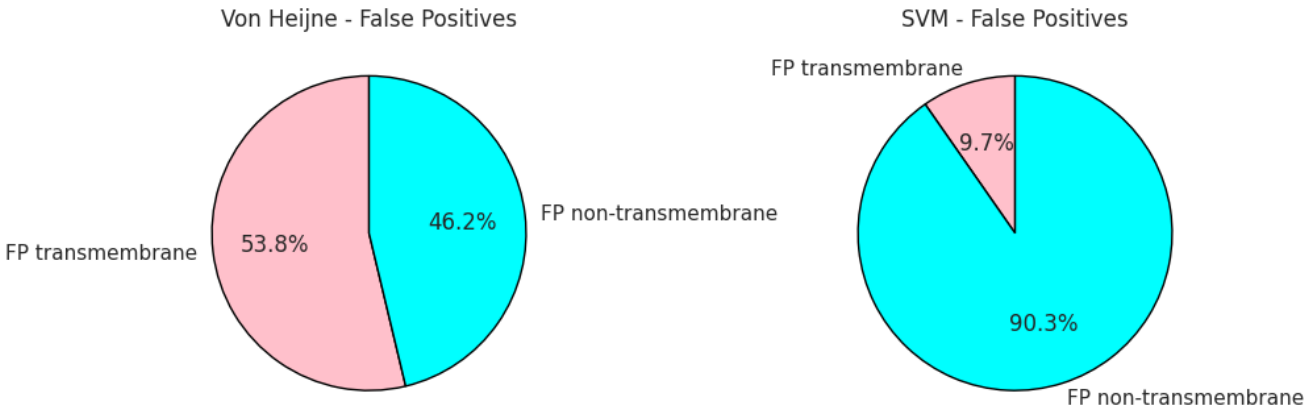

Figure 22: Distribution of false positives (FPs) for the von Heijne model and the SVM classifier, grouped by the presence of predicted transmembrane helices. The PSWM model misclassifies hydrophobic TM helices as SPs (53.8%), whereas the SVM reduces this bias substantially (FP TM = 9.7%).

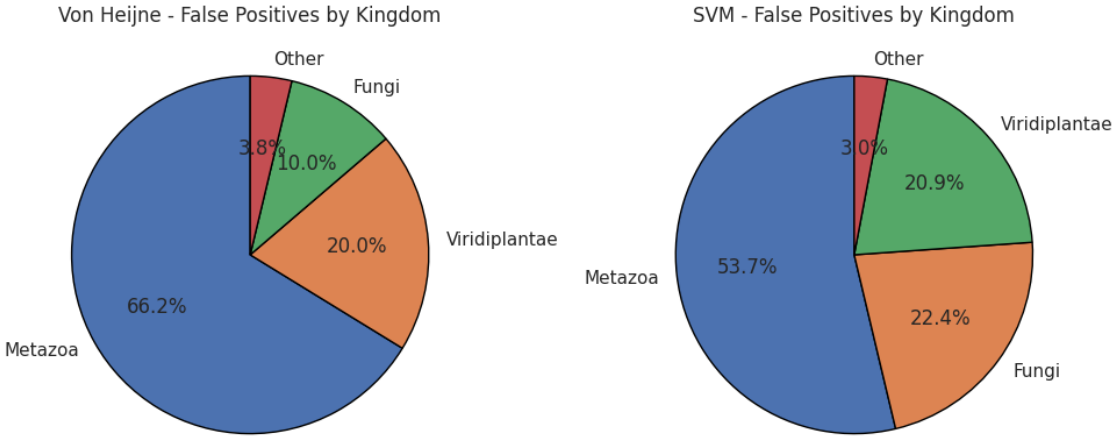

Figure 23: Taxonomic distribution of false positives for von Heijne and SVM. The PSWM model produces most errors in Metazoa (66.2%), consistent with hydrophobic region misclassification, while SVM errors are more evenly distributed, though fungal sequences remain challenging.

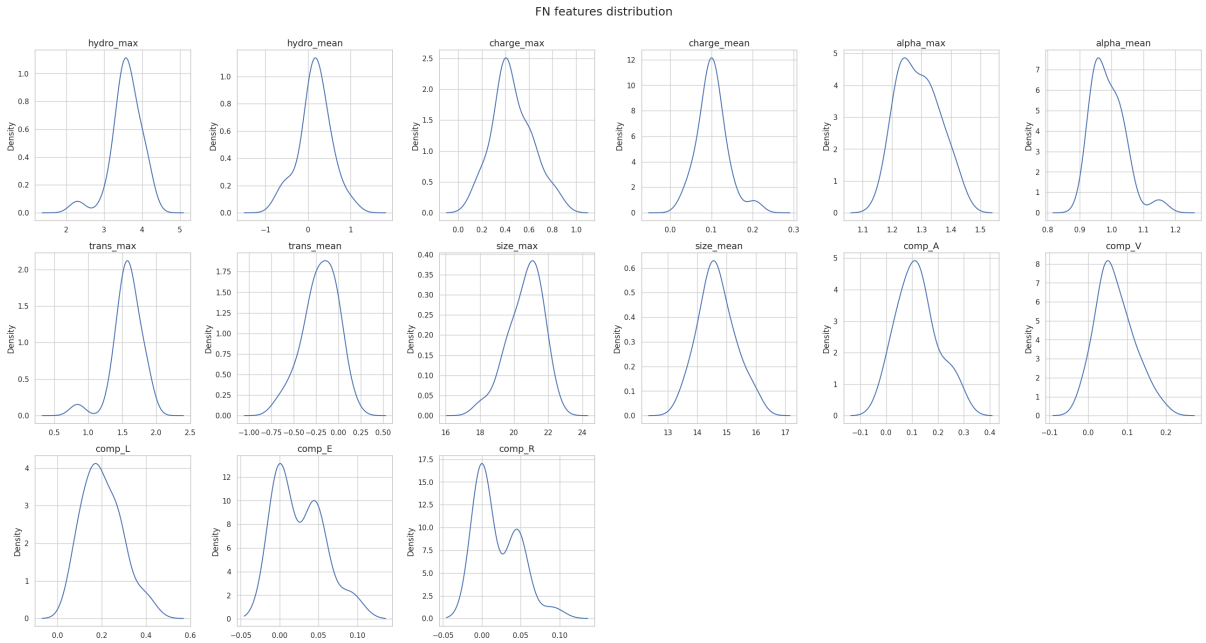

Figure 24: Kernel density estimates of engineered features for false negative (FN) sequences. FNs show reduced hydrophobicity, increased polarity, and in some cases elevated transmembrane propensity, reflecting borderline SP/TM hybrid characteristics.

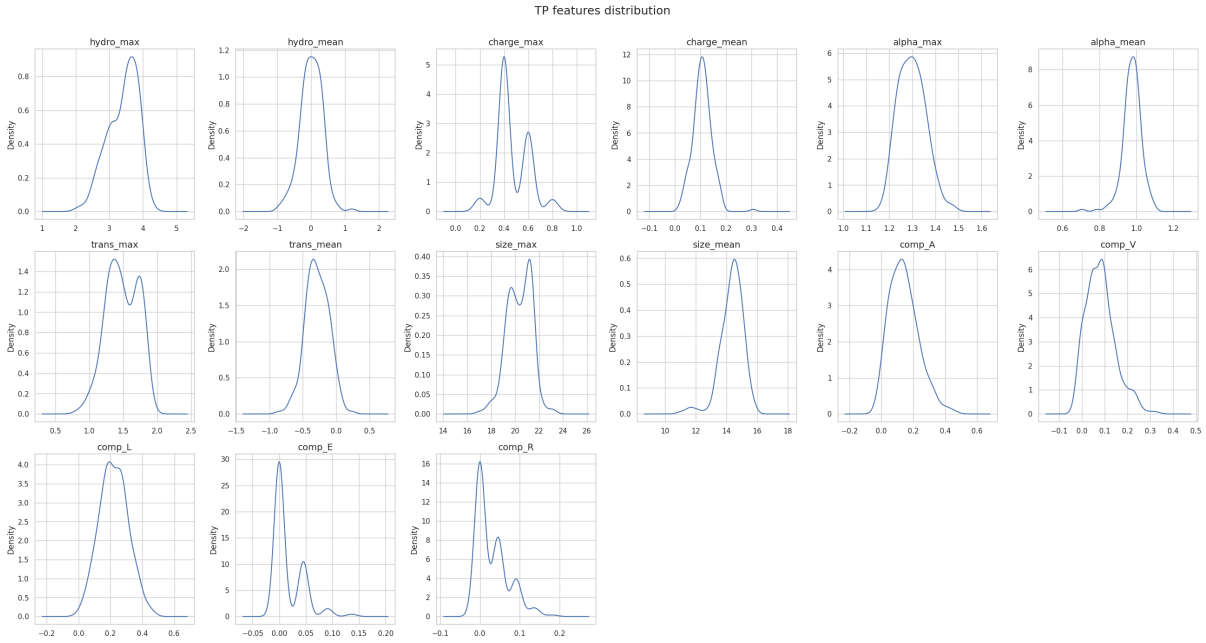

Figure 25: Feature distributions for true positives (TP). Strong hydrophobic H-regions, characteristic amino acid composition, and stable size/TM profiles clearly differentiate TP sequences from FNs.

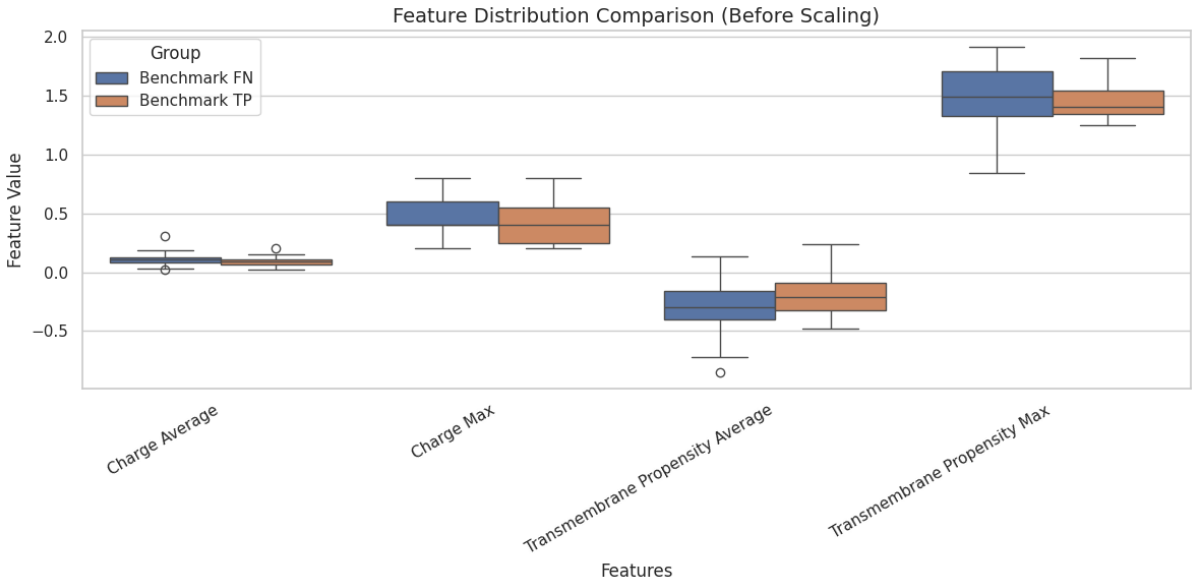

Figure 26: Feature distribution comparison between True Positives (TP) and False Negatives (FN) in the SVM model. The boxplots show the distributions of four key physicochemical features before scaling: Charge Average, Charge Max, Transmembrane Propensity Average, and Transmembrane Propensity Max. TP sequences tend to exhibit moderate charge and lower transmembrane propensity, while FN sequences show higher charge variability and an increased tendency toward transmembrane-like profiles.

5 Additional References

- Waskom, M. L. (2021). *seaborn: statistical data visualization.* Journal of Open Source Software, 6(60), 3021. https://doi.org/10.21105/joss.03021
- Hunter, J. D. (2007). *Matplotlib: A 2D graphics environment.* Computing in Science Engineering, 9(3), 90–95. https://doi.org/10.1109/MCSE.2007.55