Bioinformatics
doi: 10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: 20 November 2025
Comparative Study

Computational Biology

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

Alessia Corica[1,*], Anna Rossi[1,*], Martina Castellucci[1,*] and Sofia Natale[1,*]
[1]University of Bologna.

Correspondence: alessia.corica@studio.unibo.it, anna.rossi18@studio.unibo.it, martina.castellucci@studio.unibo.it, sofia.natale@studio.unibo.it

Associate Editor: Castrense Savojardo.

Abstract
Motivation: Signal peptides (SPs) direct proteins into the secretory pathway, making their computational identification essential for large-scale proteome annotation. Because experimental validation is accurate but slow, reliable predictive models are required. This study constructed high-quality, non-redundant eukaryotic datasets from UniProtKB and evaluated three distinct SP prediction strategies: the von Heijne position-specific scoring model, a Support Vector Machine (SVM) based on physicochemical features, and a Multilayer Perceptron (MLP) using ESM-2 embeddings.
Results: On the independent benchmark set, the von Heijne method achieved 93% accuracy (MCC = 0.656), showing high recall but an elevated false-positive rate. The SVM reached 92% accuracy (MCC = 0.690), with improved sensitivity but reduced precision for atypical or weakly hydrophobic SPs. The MLP achieved the highest performance, achieving 99% accuracy (MCC= 0.948). Misclassifications in heuristic and feature-based models were primarily associated with weakened hydrophobic cores or degraded `[A,V]XA` cleavage motifs.
Supplementary information: For further details, see the Supplementary Materials document.
Github repository:`https://github.com/Martinaa1408/LB2_project_Group_5`

1. Introduction

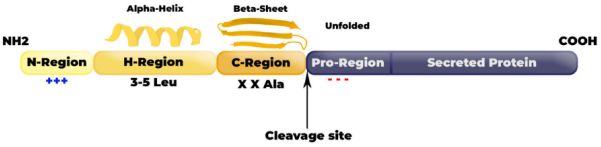
1.1 Signal peptides (SPs)

Signal peptides (SPs) are short peptides located at the N- terminus of secretory proteins. They function as a molecular "zip code" that directs proteins into the secretory pathway, the cellular route that transports newly synthesized proteins from the site of translation to their final destination, such as the plasma membrane, lysosomes, or the extracellular space.

A typical signal peptide (SP) consists of approximately 25–30 amino acids (von Heijne, 1990), although longer SPs, up to about 140 residues, are commonly found in eukaryotic proteins. Conversely, some SPs can be as short as 16 residues (Kapp and Schürmann, 2013). The general architecture of an SP comprises three distinct regions: the N-region, a positively charged segment located at the N-terminus; the H-region, a central hydrophobic core; and the C-region, which contains the cleavage site recognized by signal peptidase. Although this three-way structure is conserved across organisms, signal peptides exhibit notable variability in amino acid composition. This variability contributes to their functional flexibility in mediating protein translocation and enables their efficient operation across evolutionarily distant species.

Correct localization is essential for protein function, and errors in this process are often associated with pathological conditions such as cancer, neurodegenerative diseases, and infections (Zhang et al., 2025).

Figure 1: Signal peptide structure



In this project, three models were developed for SP prediction based on different computational strategies:

1. **Motif-Based Approaches**: Historically, methods like the Von Heijne modeled the conserved structural features of the N-, H-, and C-regions to predict cleavage sites.

2. **Machine Learning Classifiers**: Modern prediction often uses algorithms like the Support Vector Machine (SVM), which effectively analyze high-dimensional data derived from amino acid proper-

ties and sequence contexts.

3. **Deep Learning**: Advanced approaches, such as the Multilayer Perceptron (MLP), are employed to identify complex, non-linear sequence patterns for highly accurate predictions across diverse species.

1.2 Aim

The correct identification of SPs is a fundamental step in bioinformatics and molecular biology, as these short N-terminal motifs determine whether a protein will enter the secretory pathway. Understanding the presence and location of SPs provides valuable insight into the functional role of proteins and their subcellular localization.
Experimental validation through wet-lab techniques is precise but time-consuming and resource-intensive. For this reason, computational methods are increasingly adopted, offering scalability and reproducibility.

2. Materials and Methods

2.1 Data Collection

Protein datasets were retrieved from the UniProt Knowledgebase (UniProtKB (Consortium, 2023)) via its RESTful API, using the following filters to obtain high-quality eukaryotic sequences:

Table 1: Filters applied to positive and negative datasets during UniProtKB retrieval.

Criterion	Positive dataset	Negative dataset
Database	UniProtKB Swiss-sProt	UniProtKB Swiss-sProt
Taxonomy	Eukaryota (2759)	Eukaryota (2759)
Fragments	Excluded	Excluded
Sequence length	$\geq$ 40 aa	$\geq$ 40 aa
Signal peptide length	$>$ 14 aa	–
Evidence type	ECO:0000269	–
Cleavage site annotation	Required	Absent
Cellular localization	Not filtered	Non-secretory compartments
TM helices	Allowed	N-terminal TM helices tracked separately
Final dataset size	2,932 proteins	20,615 proteins (1,384 TM)

The total count of positive entries decreased from 2,949 to 2,932 because the Python script strictly selected entries that met two criteria: they must have a clearly defined cleavage site (no feature description), and the signal peptide length must exceed 14 amino acids. The negative set includes a total of 20,615 sequences, including 1,384 with N-terminal transmembrane helices—tracked separately due to their hydrophobicity potentially mimicking signal peptides.
Resulting sets were downloaded as JSON and converted into FASTA (`positive.fasta`, `negative.fasta`) and TSV formats, retaining metadata such as UniProt accession, organism, kingdom, sequence length, and eventually cleavage positions or N-terminal transmembrane segments.

2.2 Data Preparation

A structured pipeline was used to refine the datasets and ensure statistical independence. Redundancy reduction was performed separately on positive and negative sequences using `MMseqs2` (Steinegger and Söding (2017)) with parameters:

```
--min-seq-id 0.3 -c 0.4 --cov-mode 0

--cluster-mode 1
```

producing clusters with $\leq 30\%$ pairwise identity. One representative sequence per cluster was retained, and metadata were reattached via `get_tsv.py` to ensure clean one-to-one annotation.
The resulting non-redundant sequences were then partitioned using `get_sets.py` into an 80% training set and a 20% benchmark set, preserving class proportions and using `random_state = 42` for reproducibility. The training set was further split into five folds for cross-validation.

Table 2: Processing steps from raw sequences to training and benchmark sets.

Stage / Subset	Positive	Negative	Total
Initial datasets (UniProt)	2,949	20,615	23,564
After MMseqs2 clustering	1,093	8,934	10,027
Training set (80%)	874	7,147	8,021
Benchmark set (20%)	219	1,787	2,006

The obtained output files were: `pos_train.tsv`, `neg_train.tsv`, `pos_bench.tsv`, `neg_bench.tsv`. Each sequence in the training set is also annotated with a cross-validation fold index (1–5).

2.3 Data Analysis

A comprehensive set of analyses was performed to assess the biological coherence and statistical soundness of the curated datasets. Training and benchmark datasets were examined independently to verify their consistency, ensuring that expected biological patterns were maintained, potential biases were minimized, and taxonomic diversity remained sufficient to support the development of robust and generalizable predictive models.

- **Protein length distribution:** Histograms and box plots compared the lengths of proteins with and without signal peptides, allowing detection of systematic differences that could influence model performance or feature scaling.

- **Signal peptide length:** Histogram-based analyses confirmed that annotated SP lengths fell within biologically plausible ranges and helped identify potential outliers.
- **Amino acid composition:** Relative amino acid frequencies within signal peptides were computed and contrasted with global SwissProt frequencies. Bar plots highlighted characteristic trends—such as hydrophobic residue enrichment—likely to inform classification.
- **Taxonomic diversity:** Taxonomic lineages were assigned to each protein and distributions across eukaryotic kingdoms and species were visualized using bar and pie charts.
- **Cleavage site motifs:** The aligned 16-residue sequence windows (−13 to +2) surrounding the cleavage sites were processed using a tool like WebLogo (Crooks et al. (2004)) to generate a sequence logo.

2.4 Von Heijne Method

Sixteen–residue windows (−13 to +2 around the cleavage site) were extracted from experimentally validated signal peptides and used to build a Position-Specific Weight Matrix (PSWM). Amino acid frequencies at each position were computed with pseudocounts of +1 and normalized using SwissProt background frequencies. Log-odds weights were defined as

$$W_{k,j} = \log\left(\frac{M_{k,j}}{b_k}\right),$$

where $M_{k,j}$ is the observed frequency of amino acid k at position j and b_k its background frequency. Protein sequences were scored using a sliding window of length 16; for each position i:

$$S(i) = \sum_{j=1}^{16} W_{x_{i+j-1},j}.$$

The maximum score in the N-terminal region of each sequence was retained. A decision threshold was selected on a validation split by maximizing the F1-score and then fixed for all evaluations.

2.5 Support Vector Machine Method

2.5.1 Feature Extraction

The process utilizes a physicochemical encoding strategy to convert N-terminal protein sequences (residues 1–40, symmetrically extended) into a final vector of 29 numerical attributes. Feature computation relies on property scales sourced from ProtScale Gasteiger et al. (2005) and AAindex, calculating descriptors across five key categories:

- Hydropathicity (Kyte-Doolittle (Kyte and Doolittle, 1982))
- Charge distribution
- α-helix propensity (Chou-Fasman (Chou and Fasman, 1978))
- Transmembrane tendency
- Size/volume parameters

Mean and maximum values were extracted for these properties using a sliding 5-residue window and amino acid composition was also computed for the first 22 residues. The final feature vector comprised 29 attributes per sequence and was saved as `ML_features.tsv`.

2.5.2 SVM Implementation

A Support Vector Machine (SVM) was implemented to classify SP and non-SP sequences using numerical features. The model was trained under a soft–margin formulation and optimized using an RBF kernel to account for non-linear class boundaries. The RBF kernel is defined as

$$K(x_i, x_j) = \exp\left(-\gamma \|x_i - x_j\|^2\right),$$

where γ controls the influence radius of each training point. Hyperparameters were tuned via cross-validated grid search to maximize generalization performance.
Key Optimization Steps:

1. Feature Selection: Feature importance, assessed via Random Forest Gini Importance, optimized the feature set from 29 to an effective subset of 15. This process prioritized features confirming the biological properties of SP, specifically:
   - Leucine composition (comp_L).
   - α-helix mean propensity.
   - Charge mean (reflecting the N-region).
2. Hyperparameter Tuning: A Grid Search with 5-fold Cross-Validation (CV) determined the optimal model parameters to maximize the $\mathbf{F1-score}$.
   - Optimal Kernel: RBF (to model non-linearity).
   - Optimal C: 10 (Regularization strength).
   - Optimal γ: $'scale'$ (Kernel Coefficient).

2.6 Evaluation and Comparison

Stability and performance were assessed using 5-fold cross-validation, with the final metrics directly compared against a specified Benchmark performance column. The final evaluation file was saved as `vonHeijne_final.tsv`.
The Support Vector Machine (RBF kernel) was evaluated at two levels: internal evaluation on the full training/validation dataset (8021 proteins) and external benchmarking on an independent test set (2006 proteins). The final evaluation file was saved as `svm_final.tsv`.

Standard metrics included:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\text{Precision} = \frac{TP}{TP + FP} \qquad \text{Recall (TPR)} = \frac{TP}{TP + FN}$$

$$\text{FPR} = \frac{FP}{FP + TN} \qquad \text{FNR} = \frac{FN}{FN + TP}$$

$$\text{F1-score} = \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

The aim of this analysis is to quantitatively and qualitatively assess the predictive behavior of the Von Heijne model and the SVM classifier on an independent benchmark dataset, focusing on:

- Classification metrics and confusion matrices
- False positive (FP) and false negative (FN) trends
- Cleavage-site motif conservation
- Amino acid composition and signal peptide length
- Feature-level deviations (hydrophobicity, charge, TM propensity)

2.7 Multilayer Perceptron Method

The Multi-Layer Perceptron (MLP) employed a Transfer Learning approach for final sequence classification. The model leveraged 1280-dimensional sequence embeddings derived from the pre-trained ESM-2 (T33_650M) Protein Language Model, generated via Average Pooling over the last encoder layer. The MLP consisted of three fully-connected hidden layers (dimension 40 each) using *ReLU* activation, with Dropout ($p = 0.25$) applied after each layer to prevent overfitting.
Training utilized the `Adam` optimizer and `Cross-Entropy Loss`. `Optuna` was used for hyperparameter tuning to maximize the Matthews Correlation Coefficient (MCC) on the validation set. Robustness was ensured through Early Stopping (patience 10 epochs), selecting the model state with the best validation MCC.
The final evaluation file was saved as `DL_bench_res.txt`

3. Results and Discussion

The performance of the von Heijne model, SVM classifier, and MLP was assessed through cross-validation, feature analyses, benchmark evaluation, and a comparative examination of their error patterns.

3.1 Comparative Performance Evaluation

The three models were evaluated using distinct data–splitting strategies. All datasets were first divided into an 80/20 train–benchmark split. For the von Heijne model, the training portion was further partitioned into a 3:1:1 ratio, corresponding to train, validation, and an internal test set used to compute the metrics reported in Table 3.
In contrast, the SVM evaluation was based on a simple train–test split performed through `train_test_split` from scikit-learn, using the benchmark portion as the independent test set. For the MLP, the training data were instead divided into five folds following a 3:2 (train:validation) scheme, and model selection relied on the validation fold performance; the final metrics in Table 3 correspond to this validation set.
Overall, the von Heijne method displays moderate and stable performance, the SVM improves sensitivity but lowers precision, and the MLP achieves the highest accuracy, sensitivity, and MCC among the three. These differences highlight the progressive modelling capacity from motif-based rules to handcrafted features and finally deep embedding–based representations.

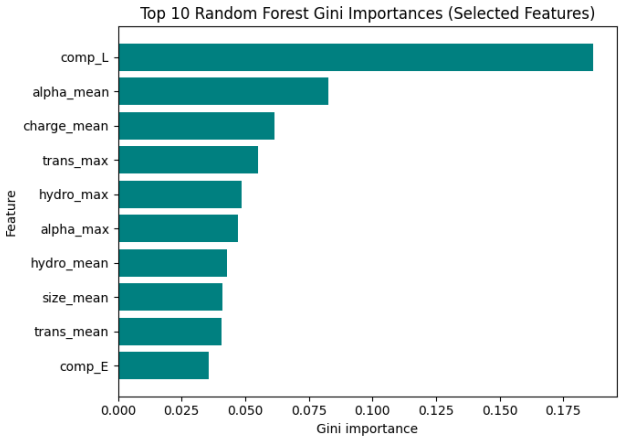
Table 3: Summary of performance evaluation metrics.

Model	Accuracy	Sensitivity	Precision	MCC
von Heijne	0.939 ± 0.002	0.756 ± 0.032	0.708 ± 0.017	0.697 ± 0.013
SVM (RBF)	0.927	0.857	0.620	0.691
MLP (ESM-2)	0.995	0.970	0.987	0.978

3.2 Feature Importance Analysis

The Random Forest Gini analysis (Figure 2) highlights a clear dominance of a few key physicochemical descriptors. Leucine composition (`comp_L`) is the most informative feature, reflecting the strong dependence of signal peptides on a well-defined hydrophobic core. α-helix mean propensity and charge-related metrics follow, capturing the structural transition between the positively charged N-region and the helix-forming H-region. Hydrophobicity peaks and transmembrane-tendency features provide secondary contributions, whereas size-based descriptors have minimal impact. Overall, the importance profile aligns with the canonical N-, H-, and C-region organization and confirms that hydrophobic enrichment and helix propensity are the primary drivers of SVM discrimination.

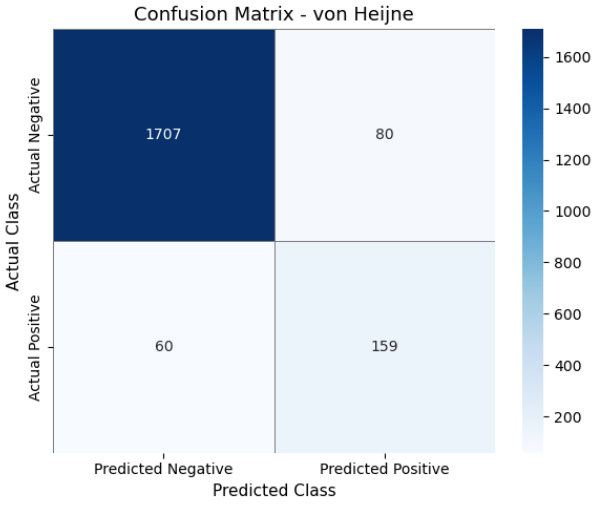
Figure 2: Feature Importance in the SVM Model



3.3 Comparative Analysis of Confusion Matrices

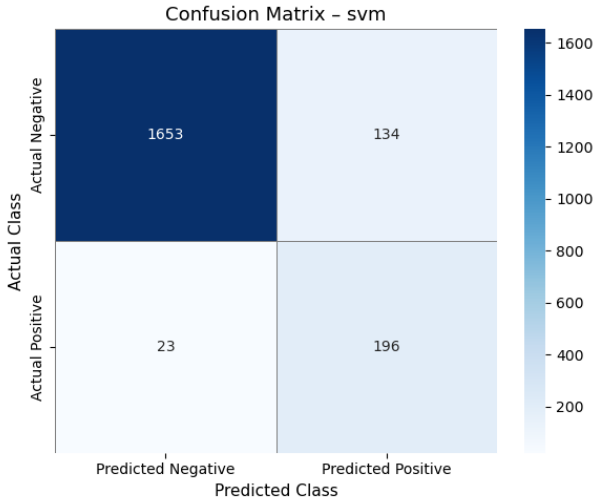
The three models exhibit distinct classification behaviours on the independent benchmark set.
The von Heijne method (Figure 3) shows high sensitivity but limited specificity, correctly detecting most signal peptides (TP = 159) while generating a substantial number of false positives (FP = 80). This pattern reflects its reliance on fixed positional motifs and hydrophobicity thresholds, which tend to overpredict borderline N-terminal regions.

Figure 3: Confusion matrix of von Heijne



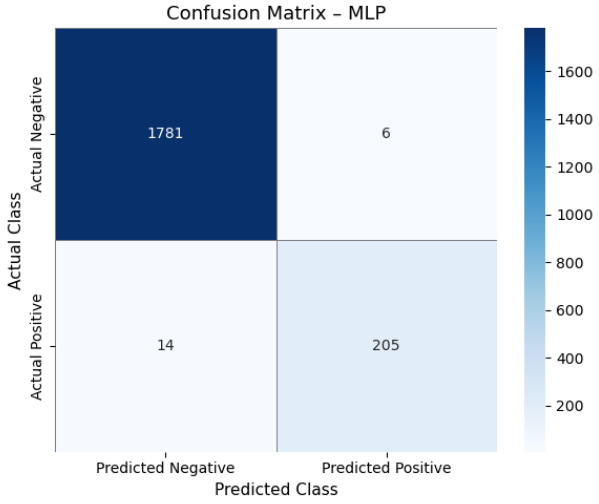
The SVM classifier (Figure 4) displays a more balanced decision boundary, reducing false negatives (FN = 23) while maintaining a moderate false-positive rate (FP = 134). Its ability to integrate physicochemical properties yields improved discrimination, although sequences with atypical hydrophobic or charge profiles still pose challenges.

Figure 4: Confusion matrix of svm



The MLP (Figure 5) achieves the best overall performance, with a nearly symmetric error distribution and minimal misclassifications (TP = 205, TN = 1781, FP = 6, FN = 14). Leveraging ESM-2 embeddings, the model captures higher-order sequence patterns that surpass both motif-based and handcrafted feature approaches.

Figure 5: Confusion matrix of MLP



3.4 Benchmark Results Comparison

The benchmark evaluation highlights clear performance differences among the three classifiers. The von Heijne model achieves solid accuracy but remains limited by its fixed scoring scheme, resulting in moderate precision and a comparatively lower MCC. The SVM improves recall substantially, detecting a larger fraction of true signal peptides, although its precision drops on the more diverse benchmark distribution, reflecting a tendency to overpredict positives. The MLP clearly outperforms both classical methods, achieving near-perfect accuracy and an MCC of 0.948, with balanced precision and recall. These results confirm the advantage of learned representations over handcrafted features or motif-based rules when generalizing to unseen eukaryotic sequences.

Table 4: Performance of the three models on the benchmark set.

Model	Acc	Prec	Rec	F1	MCC
von Heijne	0.930	0.665	0.726	0.694	0.656
SVM (RBF)	0.922	0.594	0.895	0.714	0.690
MLP (ESM-2)	0.990	0.971	0.936	0.953	0.948

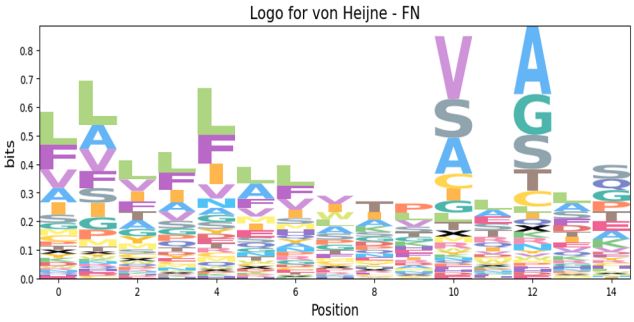
3.5 Evaluation and Comparisons

The von Heijne model and the SVM were compared by analysing their error patterns and the sequence properties underlying correct and incorrect predictions. Differences in motif strength, amino acid composition, and signal peptide length reveal the sensitivity of the von Heijne method to hydrophobicity loss and motif degradation, while the SVM shows more stable discrimination through integrated physicochemical features.

3.5.1 Motif Consistency and Length Distribution

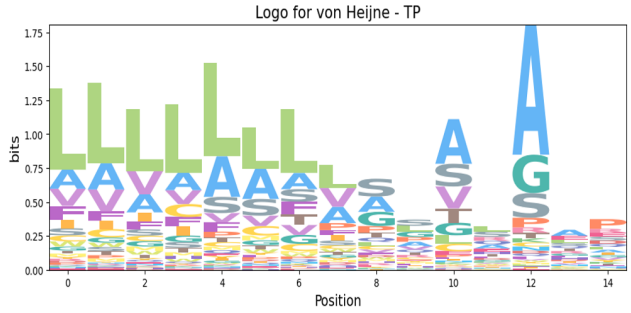
Figures 6 and 7 show the sequence logos around the cleavage sites for von Heijne false negatives (FN) and true positives (TP). False negatives display a weakened `[A,V]XA` motif and reduced upstream hydrophobicity, reflecting the loss of the canonical H-to-C region transition.

Figure 6: Sequence logo von Heijne False Negative



In contrast, true positives retain a well-defined hydrophobic core and cleavage motif consistent with (von Heijne's rule sistema). Composition analysis confirms that FNs are enriched in polar or charged residues and depleted in hydrophobic ones, suggesting that misclassifications arise from minor compositional shifts rather than annotation errors.

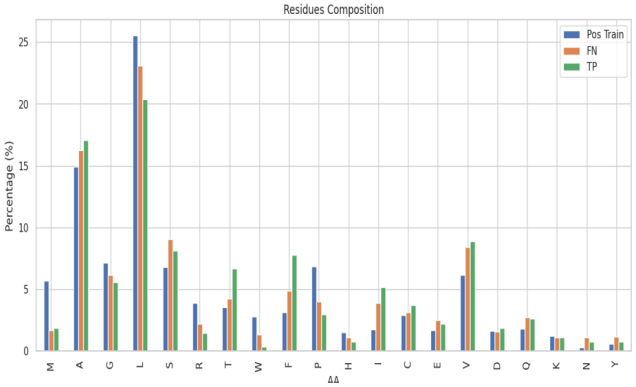
Figure 7: Sequence logo von Heijne True Positive



3.5.2 Amino Acid Frequency Distributions

The comparison of residue frequencies (Figure 8) between true positives and false negatives reveals consistent physicochemical trends. True positives display higher frequencies of L, A, and V, which define the hydrophobic H-region crucial for membrane translocation, whereas false negatives show enrichment in small polar residues such as S and T. This shift toward a less hydrophobic composition weakens cleavage recognition by the position-weight matrix, confirming that deviations in residue content are a primary driver of model inaccuracies.

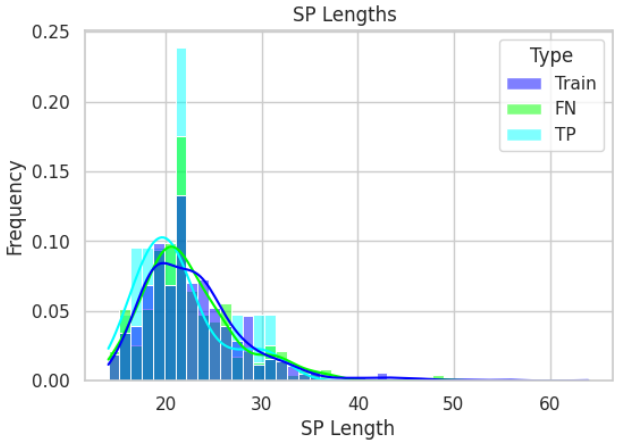
Figure 8: Amino Acid Composition of Training Set, True Positives, and False Negatives



3.5.3 Signal Peptide Length Profile

Signal peptide lengths cluster around 22 residues in both datasets, typical of eukaryotic SPs (Figure 9). Misclassified sequences tend to be slightly shorter (<18 aa), reflecting incomplete hydrophobic cores that weaken cleavage recognition. A small fraction of longer SPs (25–30 aa), mostly fungal or plant, represent lineage-specific extensions of the H-region. These trends indicate that SVM errors primarily arise from compositional and structural variability, not from dataset bias.

Figure 9: Length Profiles of True Positives and False Negatives Compared to Training SPs



3.6 MLP Training Curve Analysis

Figure 10 shows the evolution of training and validation loss across epochs for the MLP trained on ESM-2 embeddings. Both curves drop sharply during the initial iterations, indicating fast acquisition of the core decision boundary, and then converge to a stable plateau around a comparable loss value. The tight coupling between the two curves throughout training demonstrates that the model does not overfit: after the early drop, training and validation loss remain nearly indistinguishable, with only small fluctuations in the validation curve—consistent with stochastic minibatch variation rather than instability. This behavior confirms that the chosen architecture, together with dropout regularization and early stopping, maintains stable generalization while exploiting the high-dimensional ESM-2 representation.

Figure 10: Training and validation loss over epochs

3.7 Discussion

The analyses performed across motif-based, feature-based, and deep learning approaches highlight a coherent hierarchy in the capacity to capture the determinants of signal peptide identity. Biological characterisation confirms the expected organisation of SPs, with enrichment in hydrophobic residues (L, A, V, M), a positively charged N-region, and conservation of the `[A,V]XA` cleavage motif. These properties were preserved in the true positives of all models, validating both dataset construction and downstream evaluation.
The von Heijne model performs well on canonical patterns but shows limited flexibility. Its false negatives were enriched in polar residues (S, T, D, E) and tend to be shorter (<18 aa), consistent with weakened hydrophobic cores and degraded cleavage motifs. False positives often display partial hydrophobic stretches without a proper C-region transition. These findings confirm that PSWM-based scoring is highly sensitive to local compositional deviations from the idealised N–H–C framework.
The SVM improves discrimination by integrating 29 physicochemical features. Feature importance analysis identifies leucine composition, α-helix propensity, and charge as the dominant descriptors, aligning with the biological roles of the N- and H-regions. Its confusion matrix shows a more balanced distribution of FP and FN than the von Heijne model, although sequences with atypical hydrophobicity profiles still challenge the classifier. This indicates that handcrafted features, while informative, cannot fully capture the subtle sequence variability present in eukaryotic SPs.
The MLP achieves the strongest performance, supported by near-perfect benchmark metrics and symmetric error rates (FP = 6, FN = 14). The training and validation curves demonstrate stable convergence without overfitting. Leveraging ESM-2 embeddings, the model effectively encodes long-range contextual information and lineage-specific variations, allowing it to recognise both canonical SPs and non-standard variants with reduced hydrophobicity or weakened motifs. This capacity surpasses both the rigidity of the von Heijne model and the limited expressiveness of engineered features in the SVM.
Taken together, the results reveal a clear progression in modelling power: rigid motif scoring is surpassed by feature-based learning, which is in turn outperformed by deep contextual embeddings. The findings underscore the importance of continuous, high-dimensional representations for accurately modelling the biophysical and evolutionary diversity of signal peptides.

3.7.1 Limitations

Despite the robustness of the analyses and the strong performance of the MLP model, several limitations must be acknowledged. First, the study focuses exclusively on eukaryotic proteins with experimentally validated annotations, which may limit generalisation to prokaryotic or poorly annotated taxa. The negative dataset includes sequences with N-terminal transmembrane helices, but other borderline classes—such as signal anchors or low-complexity secretory regions—were excluded, potentially underestimating real-world ambiguity. The SVM relies on handcrafted physicochemical features, which, although interpretable, cannot fully capture long-range dependencies or evolutionary signals. For the MLP, performance is strongly influenced by the choice of pretrained ESM-2 embeddings; alternative architectures or larger models might further improve generalization. Finally, benchmarking was performed on held-out UniProtKB sequences, and real deployment settings may include noisier or incomplete annotations. To increase robustness, cross-validation should also be applied to the SVM and MLP models, as done for the von Heijne method. These limitations highlight the need for broader datasets, cross-species evaluation, and exploration of alternative deep learning architectures.

4. Conclusions

This study compared three distinct strategies for signal peptide prediction using rigorously curated, non-redundant eukaryotic datasets. The results reveal a clear progression in modelling effectiveness: the von Heijne method reliably identifies canonical motifs but is limited by sensitivity to hydrophobicity loss and motif

weakening; the SVM improves discrimination through physicochemical descriptors yet remains constrained by the expressiveness of handcrafted features. In contrast, the MLP leveraging ESM-2 embeddings achieves near-perfect benchmark performance, capturing both canonical and compositionally atypical SPs with balanced error rates and stable generalisation.
Overall, these findings demonstrate that deep, context-aware representations provide substantial advantages for SP classification, surpassing both motif-driven and feature-based approaches. The study underscores the value of modern protein language models as a robust foundation for accurate and scalable signal peptide annotation across diverse eukaryotic proteomes.

References

P. Y. Chou and G. D. Fasman. Prediction of the secondary structure of proteins from their amino acid sequence. *Biochemistry*, 13(2):222–245, 1978. doi: 10.1021/bi00699a002.

The UniProt Consortium. Uniprot: the universal protein knowledgebase in 2023. *Nucleic Acids Research*, 51(D1):D523–D531, 2023. doi: 10.1093/nar/gkac1052.

Gavin E Crooks, Gary Hon, John-Marc Chandonia, and Steven E Brenner. Weblogo: a sequence logo generator. *Genome Research*, 14(6):1188–1190, 2004. doi: 10.1101/gr.849004.

E. Gasteiger, C. Hoogland, A. Gattiker, S. Duvaud, M. R. Wilkins, R. D. Appel, and A. Bairoch. Protein identification and analysis tools on the expasy server. In John M. Walker, editor, *The Proteomics Protocols Handbook*, pages 571–607. Humana Press, 2005.

Katja Kapp and Asngar Schürmann. Post-targeting functions of signal peptides. *Protein transport into the endoplasmic reticulum*, pages 1–16, 2013.

J. Kyte and R. F. Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132, 1982. doi: 10.1016/0022-2836(82)90515-0.

Martin Steinegger and Johannes Söding. Mmseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, 35(11):1026–1028, 2017. doi: 10.1038/nbt.3988.

G. von Heijne. The signal peptide. *Journal of Membrane Biology*, 115:195–201, 1990.

S. Zhang, Z. He, H. Wang, and J. Zhai. Signal peptides: From molecular mechanisms to applications in protein and vaccine engineering. *Biomolecules*, 15 (6):897, 2025.