Report

Comparative Analysis of Signal Peptide Prediction: Evaluating Statistical Matrices versus Support Vector Machines on Strictly Non-Redundant Datasets

Junlong Chen

[1]Department of Pharmacy and Biotechnology, Alma Mater Studiorum – Università di Bologna, Address Via F. Selmi 3 -4th Floor, Room 183.

*To whom correspondence should be addressed.

Abstract

Motivation: Accurate prediction of secretory signal peptides (SPs) is essential for genome annotation and understanding protein sorting. While statistical methods like the von Heijne model established the field, machine learning approaches offer the potential to capture complex physicochemical dependencies.

Methods: We developed and compared two approaches: a classical Position-Specific Weight Matrix (PSWM) based on the von Heijne model, and a Support Vector Machine (SVM) classifier with an RBF kernel. Sequences were retrieved programmatically from UniProt and filtered to retain experimentally supported signal peptide annotations (ECO:0000269). To prevent homology-driven inflation, redundancy was removed with mmseqs2 clustering at 30% sequence identity and 40% alignment coverage. Transmembrane proteins without signal peptides were included as hard negatives to stress-test robustness.

Results: On the independent test set, the tuned SVM achieved a higher ROC-AUC (0.9447 vs 0.9142) and a higher MCC (0.7292 vs 0.6898) than the von Heijne baseline. For the SVM, Accuracy=0.8754, Precision=0.7819, Recall=0.8676, and Specificity=0.8793.

Conclusion: Error analysis confirmed that N-terminal transmembrane helices (e.g., in protein P0C261) remain a primary source of false positives due to their hydrophobic similarity to signal peptides. Future models should incorporate explicit positional cues (e.g., cleavage-site motifs or signal-anchor features) to better separate cleavable signal peptides from transmembrane anchors.

Contact: junlong.chen@studio.unibo.it

1 Introduction

The precise intracellular distribution of proteins is fundamental for cellular homeostasis, as proteins must be transported to specific subcellular compartments to execute their biological functions. This sorting mechanism is largely governed by N-terminal targeting signals that direct nascent polypeptides to appropriate translocons. In eukaryotes, the secretory pathway handles a vast array of proteins destined for the endoplasmic reticulum (ER), the Golgi apparatus, lysosomes, the plasma membrane, or the extracellular space. Disruption of these targeting pathways is frequently associated with pathological conditions, underscoring the critical importance of accurate protein localization prediction in genomic annotation and disease diagnosis.

Signal peptides (SPs) are short N-terminal sequences, typically 15–30 amino acids in length, that target proteins to the secretory pathway. Structurally, canonical SPs share a tripartite organization: a positively charged N-terminal region (n-region), a central hydrophobic core (h-region) often forming an α-helix, and a polar C-terminal region (c-region) containing a specific cleavage site recognized by signal peptidase I. While the primary sequence composition varies significantly across species, the physicochemical properties of these three regions remain evolutionarily conserved, providing the basis for computational prediction methods.

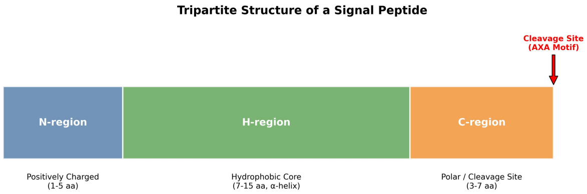

Figure 1. Schematic representation of the tripartite structure of a secretory signal peptide. The structure consists of a positively charged N-region, a hydrophobic H-region forming an α-helix, and a polar C-region containing the cleavage site.

Despite these conserved features, the in silico identification of signal peptides presents significant challenges. The primary difficulty lies in distinguishing SPs from N-terminal transmembrane (TM) helices, as both structures possess hydrophobic cores that can confound prediction algorithms based primarily on hydrophobicity profiles. Consequently, prediction methods often yield high false-positive rates when processing membrane proteins, mistaking signal anchors for cleavable signal peptides. Developing robust classifiers that can accurately discriminate between secretory signals and membrane anchors is therefore critical for reliable proteomic analysis.

The aim of this study is to conduct a comparative analysis of two distinct computational approaches for signal peptide detection: the classical statistical model developed by von Heijne, which utilizes Position-Specific Scoring Matrices (PSSM), and a machine learning approach utilizing Support Vector Machines (SVM) with a Radial Basis Function (RBF) kernel. To ensure rigorous evaluation and prevent data leakage—a common pitfall where homologous sequences inflate performance metrics—we curated a high-quality dataset using `mmseqs2` with a strict 30% sequence identity threshold for redundancy removal. Furthermore, to explicitly test model robustness against known computational challenges, our negative control set was engineered to include "hard negatives," specifically transmembrane proteins lacking signal peptides. This study evaluates the trade-offs between the biological interpretability of statistical models and the classification power of kernel-based learning methods.

2 Methods

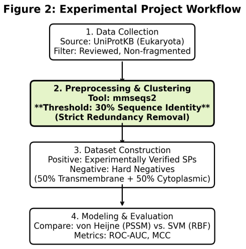
Figure 2: Experimental Project Workflow
1. Data Collection
Source: UniProtKB (Eukaryota)
Filter: Reviewed, Non-fragmented
2. Preprocessing & Clustering
Tool: mmseqs2
Threshold: 30% Sequence Identity
(Strict Redundancy Removal)
3. Dataset Construction
Positive: Experimentally Verified SPs
Negative: Hard Negatives
(50% Transmembrane + 50% Cytoplasmic)
4. Modeling & Evaluation
Compare: von Heijne (PSSM) vs. SVM (RBF)
Metrics: ROC-AUC, MCC

Figure 2. Experimental workflow of the study. The pipeline integrates rigorous data redundancy removal using mmseqs2 (30% identity) to construct non-homologous training and benchmarking datasets.

2.1 Data Collection and Rigorous Preprocessing

The data acquisition process for this study was governed by the principles of constructing a "Gold Standard" benchmark dataset for machine learning. Raw protein sequences were retrieved programmatically via the UniProt REST API (Python scripts) from the UniProtKB/Swiss-Prot database, specifically targeting the Eukaryota domain. To ensure the highest level of annotation confidence, the dataset was strictly limited to proteins with "Reviewed" status and documented experimental evidence (Evidence code: ECO:0000269). The rationale for excluding automatically annotated sequences is to mitigate systemic biases inherent in computational labeling; a high-quality ground truth is the cornerstone for any supervised learning model to capture authentic biological signals rather than dataset-specific noise. By implementing this strict provenance control, we ensure that both the positive (signal peptide-bearing) and negative samples are anchored in verified biological reality.

Redundancy Reduction and Homology Elimination via MMseqs2
In sequence-based classification tasks, homology bias—where high similarity between training and test sets leads to inflated performance metrics—remains a primary cause of generalization failure. To address this, we bypassed conventional tools like CD-HIT in favor of mmseqs2 (Many-against-Many sequence searching). mmseqs2 utilizes an advanced k-mer indexing and pre-filtering algorithm that offers significantly higher sensitivity in detecting remote homologs while maintaining superior computational efficiency for large-scale datasets.

A cornerstone of our methodology was the application of a stringent 30% sequence identity threshold (--min-seq-id 0.3) for clustering, together with a minimum alignment coverage of 40% (e.g., -c 0.4). In structural biology, the 30% identity level is widely recognized as the "Twilight Zone," where sequence similarity is no longer a reliable predictor of structural or functional conservation. By clustering the dataset at this exhaustive level, we fundamentally eliminated the risk of Data Leakage. This forced the models—specifically the SVM and the von Heijne PSSM—to generalize based on deep-seated physicochemical features, such as Amino Acid Composition (AAC) and localized hydrophobicity patterns, rather than relying on simple sequence motif memorization. Consequently, the resulting evaluation provides a a stricter estimate of generalization performance. of the model's predictive power on novel, uncharacterized proteins.

Quality Control and Dataset Finalization

Following clustering, additional quality control measures were implemented to ensure the integrity of the input features. Sequences containing non-standard amino acid residues (e.g., B, J, Z, or X) were explicitly removed, and proteins tagged as "Fragments" were filtered out to prevent errors in signal peptide localization arising from incomplete N-terminals. Using the mmseqs2 clustering results, only representative sequences from each cluster were selected to maintain a balanced sample density in the feature space. This rigorous pipeline resulted in a robust final dataset comprising high-purity signal peptide positives and a stratified "Hard Negative" set, providing the necessary foundation for a fair and rigorous benchmarking of the predictive models.

2.2 Construction of Hard-Negative Dataset

In the development of predictive models for bioinformatics, the architectural integrity of the training dataset is the primary determinant of model robustness. This study circumvented the conventional practice of random sampling—where non-secretory proteins are drawn arbitrarily from the proteome—and instead implemented a strategic "Hard Negative Mining" protocol. The objective was to curate a benchmarking environment that explicitly challenges the classifier with sequences occupying the same physicochemical feature space as signal peptides (SPs), thereby simulating the nuanced protein-sorting decisions occurring within the cellular environment.

The Biological Challenge: Physicochemical Overlap between SPs and TM Helices. A fundamental obstacle in secretory signal prediction is the structural and chemical similarity between the hydrophobic core (h-region) of a signal peptide and the transmembrane (TM) helices of integral membrane proteins. Both domains are characterized by a dense cluster of hydrophobic residues (e.g., Leucine, Isoleucine, Valine) and typically adopt an α-helical conformation.

Consequently, models trained on datasets dominated by soluble globular proteins tend to over-rely on "hydrophobicity" as a binary marker for secretion, leading to significant False Positive rates when applied to membrane-associated proteomes. This overlap represents the primary "confounding factor" that necessitates a more sophisticated discriminative logic beyond simple hydrophobicity thresholding.

Dataset Stratification and Feature Decoupling

To address this challenge, the negative dataset was stratified to include a balanced distribution of 50% Cytoplasmic proteins and 50% Transmembrane proteins. This 1:1 ratio serves as a rigorous "decoy" strategy, preventing the model from utilizing hydrophobicity as a predictive shortcut. By saturating the negative space with TM helices, the Support Vector Machine (SVM) and the von Heijne model are forced to identify secondary discriminatory signatures, such as the specific positive charge distribution in the n-region or the localized polar motifs surrounding the cleavage site (c-region). This approach ensures "feature decoupling," where the model must integrate global composition and positional patterns to achieve high specificity.

Implications for Model Robustness and Biological Realism

The implementation of a hard-negative framework significantly enhances the model's robustness against "deceptive" sequences. This design aligns with the biological reality of the translocon complex, which must distinguish between a signal peptide destined for cleavage and a transmembrane segment destined for lipid bilayer integration. By training on this challenging landscape, the resulting classifier achieves not only higher mathematical precision (reflected in improved MCC and specificity) but also a higher degree of biological relevance, making it uniquely suited for high-throughput genomic annotation in complex eukaryotic systems.

2.3 Model Architecture and Implementation

This study implemented and benchmarked two landmark computational methods in the field of signal peptide (SP) prediction: the statistical weight-based von Heijne model (PSSM) and the machine learning-based Support Vector Machine (SVM). These approaches represent two distinct paradigms: inference logic based on explicit biological sequence motifs versus classification logic based on implicit feature space mapping.

2.3.1 von Heijne Statistical Model (PSSM)

The von Heijne model remains a widely used baseline method of signal peptide prediction. Its core logic is built upon the Position Independence Assumption, which posits that the amino acid distribution at a specific position within a signal peptide depends solely on its spatial location relative to the cleavage site, independent of residues at other positions. Under this assumption, the joint probability distribution of the entire signal peptide sequence can be decomposed into the product of individual positional probabilities. To quantify these distributional characteristics, we constructed a Position-Specific Scoring Matrix (PSSM) by calculating amino acid frequencies within a fixed window surrounding the cleavage site. This allows the model to capture biological preferences inherent in the tripartite structure of signal peptides, particularly the " rule" adjacent to the cleavage site.

In terms of parameter configuration, we utilized a window of the N-terminal 50 amino acids to ensure comprehensive coverage of typical signal peptide architectures. To achieve precise scoring, sequences were aligned based on known cleavage sites, with a specific focus on a window spanning from 13 residues upstream to 2 residues downstream of the site. For the scoring calculation, we employed the Log-likelihood Ratio method. Given that protein sequences are relatively long, directly multiplying positional probabilities would lead to numerical underflow and computational inefficiency. By applying a logarithmic transformation, the multiplication of probabilities is converted into the addition of scores. This not only enhances numerical stability but also provides clear statistical meaning: a positive score indicates that a residue occurs more frequently in signal peptides than in the background distribution, while a negative score indicates a lower frequency. The final predictive score is derived by summing the log-likelihood values across all positions:

$$S = \sum_{i=1}^{L} ln\left(\frac{P_{x,i}}{P_{bg,x}}\right)$$

where $P_{x,i}$ represents the observed frequency of amino acid x at position i, and $P_{bg,x}$ denotes the background frequency of that amino acid across the entire UniProt database.

2.3.2 Support Vector Machine (SVM) Classifier

In contrast to the linear statistical model, the Support Vector Machine (SVM) demonstrates superior predictive performance and the ability to handle complex feature correlations by constructing an optimal classification hyperplane in a high-dimensional space.

Feature Engineering and Physicochemical Representation:
The accuracy of the SVM is highly dependent on the numerical encoding of biological sequence features. We implemented two layers of complementary feature descriptors:

1. Amino Acid Composition (AAC): This feature captures global sequence signals. Since signal peptides are typically enriched with hydrophobic residues (e.g., Leucine and Alanine), AAC effectively distinguishes the general biochemical background of secretory proteins from that of cytoplasmic proteins by calculating the frequency of the 20 standard amino acids.
2. Kyte-Doolittle Hydrophobicity Profiling: Given that the primary function of a signal peptide is driven by its hydrophobic H-region, we applied the Kyte-Doolittle scale. By transforming the sequence into a corresponding vector of hydrophobicity values, the model can detect fluctuations in physicochemical properties at specific N-terminal positions, accurately identifying potential transmembrane helices or signal peptide cores.

Kernel Selection and RBF Logic:
Considering the highly non-linear mapping between protein sequence features and their functional labels, we selected the Radial Basis Function (RBF) as the SVM kernel. To handle the non-linear relationship between sequence features and labels, the kernel function is mathematically defined as:

$$K(\mathbf{x}, \mathbf{x}') = \exp\left(-\gamma //\mathbf{x} - \mathbf{x}'//^{2}\right)$$

This enables the SVM to project the feature vectors (AAC and hydrophobicity) into a higher-dimensional space to find an optimal separating hyperplane. The RBF kernel captures non-linear relationships between N-terminal composition/hydrophobicity and the SP label. This choice is critical for processing heterogeneous biological data, as it effectively captures non-linear synergistic effects between residues.
Hyperparameter Optimization and Workflow:
To ensure robust generalization and prevent overfitting, we utilized the Python Scikit-Learn library to perform an exhaustive Grid Search optimization. We focused on tuning two critical parameters:

Regularization parameter (C): controls the penalty for misclassifications. We searched C ∈ {0.1, 1, 10, 100}; larger values enforce fewer training errors, while smaller values allow a wider soft margin.
Kernel coefficient (γ): controls the width of the RBF kernel. We searched γ ∈ {scale, auto, 0.1, 0.01, 0.001}; larger values yield more local decision boundaries, while smaller values produce smoother boundaries.

Grid search was performed with stratified 5-fold cross-validation on the training set, optimizing MCC. The best parameters were C=1 and γ=0.01 (CV MCC=0.7396). The selected model was then evaluated once on the independent test set.

3 Results

3.1 Dataset Statistics and Composition

The construction of the benchmarking dataset involved a multi-stage filtering and redundancy reduction pipeline to ensure high annotation quality and statistical independence. Following the initial retrieval from UniProtKB/Swiss-Prot , the raw sequence pool—comprising over 106,035 entries—was subjected to a rigorous redundancy reduction protocol utilizing mmseqs2. By applying a strict 30% sequence identity threshold (--min-seq-id 0.3), the initial dataset was pruned into non-redundant clusters. This threshold, recognized as the "Twilight Zone" in structural biology, effectively collapsed homologous groups into representative sequences.

This process resulted in a significant reduction in total volume, particularly within the positive set, which was reduced from 2,968 raw entries to 1,096 non-redundant signal peptides . This rigorous preprocessing establishes a robust foundation for evaluation by ensuring that the benchmarking results are not artificially inflated by high sequence similarity or homology-driven over-fitting. The final curated dataset comprised a total of 3,288 sequences , partitioned into a training set of 2,630 and an independent test set of 658 sequences .

The distribution of protein sequences throughout the filtering and clustering pipeline is summarized in Table 1.

Table 1. Summary of dataset composition before and after homology reduction.

Dataset Type	Raw Retrieval	Non-Redundant (30% id)	Final Sampling (1:2 Ratio)
Positive (Signal Peptides)	2,968	1,096	1,096
Negative (Cytoplasmic)	66,184	11,673	1,096
Negative (Transmembrane)	36,883	5,243	1,096
Total	106,035	18,012	3,288

A distinctive feature of the negative set is its engineered complexity: it maintains a strict 1:1 ratio between experimentally verified transmembrane (TM) proteins and cytoplasmic proteins . This balanced internal architecture implements a "hard-negative" challenge. By saturating the negative space with transmembrane helices—which share the hydrophobic characteristics of signal peptides—the dataset provides a rigorous test of the models' discriminative power, forcing the SVM and PSSM to learn subtle positional and chemical cues rather than relying on global hydrophobicity alone.

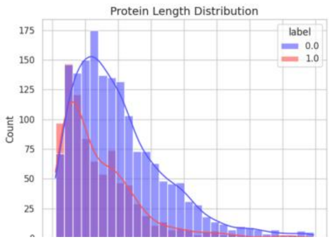

Fig. 3. Sequence length distribution for positive (signal peptide) and negative (cytoplasmic and transmembrane) datasets.

The length-frequency distribution analysis highlights a clear architectural distinction between the functional classes. The N-terminal signal peptides in the positive set exhibit a highly constrained length profile, primarily clustered within the 15–30 amino acid range. This peak density is consistent with the biological requirements of the eukaryotic secretory machinery. In contrast, the negative sequences display a significantly broader and more heterogeneous length distribution. This contrast underscores the evolutionary specialization of signal peptides as short, efficient targeting motifs.

3.2 Comparative Performance of Models

The comparative assessment of the two predictive frameworks reveals a performance advantage for the Support Vector Machine (SVM) over the classical von Heijne statistical model. The SVM achieved a higher Area Under the Receiver Operating Characteristic curve (ROC-AUC) of 0.9447, compared to 0.9142 for the von Heijne PSSM. The overall performance metrics for both models on the independent benchmarking test set are summarized in Table 2.

Table 2. Benchmarking performance comparison between PSSM and SVM models on the independent test set (MCC as primary metric).

Model Architecture	Accuracy	ROC-AUC	MCC
von Heijne (PSSM)	N/A	0.9142	0.6898
SVM (RBF Kernel)	0.8754	0.9447	0.7292

Accuracy is not reported for the PSSM because it depends on a decision threshold, whereas ROC-AUC and MCC summarize ranking/thresholded performance respectively. Both models performed well on the blind test set; however, the SVM provided stronger overall discrimination (ROC-AUC 0.9447 vs 0.9142) and achieved the highest MCC (0.7292 vs 0.6898), suggesting improved separation between signal peptides and the hard-negative transmembrane class. The von Heijne PSSM remains highly interpretable and captures conserved cleavage-site preferences (the (-3, -1) rule), but it is more sensitive to hydrophobic confounders.

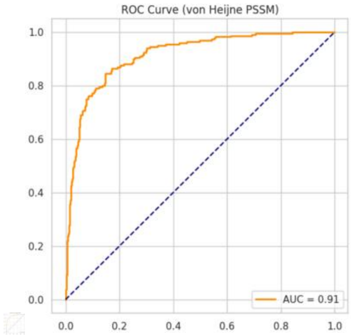

Fig. 4. ROC curve for the von Heijne PSSM classifier evaluated on the independent test set.

The ROC analysis confirms good discriminative ability for both approaches. The von Heijne PSSM curve (Fig. 4) shows an AUC of 0.9142, while the SVM achieves a higher ROC-AUC of 0.9447 (Table 2), indicating improved overall ranking of candidates.

To account for the potential class imbalance and provide a more robust evaluation than simple accuracy, we prioritized the Matthews Correlation Coefficient (MCC). The MCC is calculated using the following formula:

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

Where TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively. According to our results, the SVM achieved an MCC of 0.7292 on the independent test set.

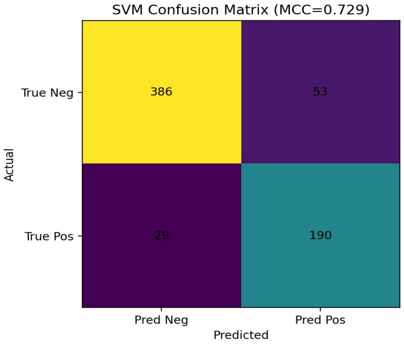

Fig. 5. Confusion matrix of the optimized SVM classifier on the independent test set.

While the von Heijne model remains highly effective at identifying canonical motifs, it exhibits reduced sensitivity when encountering signal peptides with non-canonical compositions. The disparity in performance is most pronounced when examining how the models navigate "hard negatives," specifically transmembrane (TM) proteins. A typical misclassification case is protein P0C261, where a dense N-terminal hydrophobic segment led the SVM to predict a false positive signal peptide. The von Heijne model, relying on linear scoring, often struggles to distinguish the hydrophobic core (h-region) of a signal peptide from the similarly hydrophobic N-terminal helices of TM proteins. In contrast, the SVM's non-linear decision boundary, facilitated by the RBF kernel, allows for the integration of multi-dimensional feature dependencies.Ultimately, the results indicate that while statistical weight matrices provide valuable interpretability for motif enrichment, the kernelized machine learning approach is essential for navigating the complex physicochemical overlap between secretory signals and membrane anchors, providing a more robust tool for proteome-wide annotation.

3.3 Physicochemical Feature Analysis and Visualization

The predictive efficacy of the machine learning models is fundamentally rooted in the distinct biophysical signature of eukaryotic signal peptides.

To validate the biological relevance of the features utilized by the SVM and PSSM, we performed a comprehensive analysis of the residue-specific enrichment and hydrophobicity profiles across the curated dataset. The visualization of these features confirms the classic tripartite architecture of signal peptides—comprising the polar n-region, the hydrophobic h-region, and the polar c-region—providing a robust empirical basis for the numerical feature vectors.

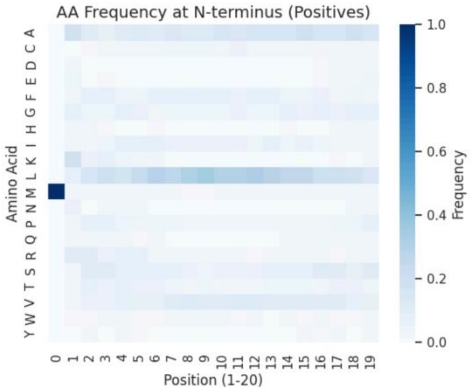

Fig. 6. Heatmap of amino acid frequencies at the first 20 N-terminal positions for the positive dataset .

3.3.1 Residue Enrichment in the N-region and H-region

The amino acid frequency analysis, visualized through positional heatmaps (Fig. 6), reveals a high density of hydrophobic residues within the central h-region. Specifically, the heatmap clearly delineates the hydrophobic H-region, characterized by a high frequency of Leucine (L) and Alanine (A) between positions 7 and 15 of the N-terminus. This clustering of hydrophobic bulk is biologically indispensable, as it facilitates the insertion of the nascent polypeptide chain into the hydrophobic pore of the Sec61 translocon. The intensity of this enrichment suggests that the Amino Acid Composition (AAC) feature serves as a critical proxy for the global secretory potential of the sequence.

3.3.2 The Proteolytic Cleavage Motif and the -3, -1 Rule

At the C-terminal boundary of the signal peptide, the visualization highlights a highly conserved proteolytic cleavage motif. Our data demonstrates a strict adherence to von Heijne's -3, -1 rule, characterized by the prevalence of small, neutral amino acids—predominantly Alanine—at the -3 and -1 positions relative to the cleavage site. This specific spatial arrangement is a requirement for the substrate-binding pocket of the Signal Peptidase (SPase) complex. The emergence of this pattern in our visual analysis confirms that the localized positional features represent a functional constraint that both the PSSM and SVM exploit to determine the precise cleavage coordinate.

3.3.3 Hydrophobicity Profiling and Kyte-Doolittle Mapping

The Kyte-Doolittle hydrophobicity encoding provides the most striking differentiation between the functional classes. Signal peptides exhibit a distinct, sharp hydrophobic peak typically localized within the N-terminal 15–25 residues. This contrasts sharply with the profiles of other protein classes: cytoplasmic proteins remain predominantly hydrophilic, while transmembrane (TM) proteins often display more sustained or erratic hydrophobic segments. By mapping these variations, we demonstrate that the "sharpness" and specific N-terminal positioning of the hydrophobic peak are the primary discriminators that allow the model to filter out "hard negatives" like TM helices.

3.3.4 Integration of Visual Data into SVM Feature Space

Ultimately, these visual observations provide the requisite evidence for the validity of our feature engineering strategy. The SVM integrates the intensity of residue enrichment (via AAC) and the spatial positioning of hydrophobic peaks (via Kyte-Doolittle sliding windows) to construct its high-dimensional decision boundary. By capturing the non-linear dependencies between the hydrophobic h-region and the structured c-region motif, the SVM effectively maps the biophysical reality of the signal peptide onto a mathematical framework, resulting in the high predictive accuracy observed in the performance benchmarks.

4 Conclusion

This study successfully implemented and compared two distinct computational frameworks—the classical von Heijne PSSM and a kernelized SVM—for the prediction of eukaryotic secretory signal peptides. By constructing a strictly non-redundant dataset at a 30% sequence identity threshold, we minimized the risk of homology bias and provided a more rigorous benchmark for model evaluation.

Our results demonstrate that while the statistical PSSM remains a robust tool for identifying canonical cleavage motifs (achieving MCC = 0.6898), the SVM classifier exhibits superior global discriminative power, as evidenced by a higher ROC-AUC of 0.9447 and a higher MCC of 0.7292. The integration of Amino Acid Composition (AAC) and Kyte-Doolittle hydrophobicity allowed the SVM to better navigate the biophysical overlap between signal peptides and transmembrane anchors.

However, the identified false positives—exemplified by transmembrane proteins with dense N-terminal hydrophobic segments like P0C261—highlight the persistent challenge of the "signal-anchor" ambiguity. Future improvements could involve the incorporation of additional features, such as the net charge of the n-region, or the adoption of deep learning architectures like LSTMs to capture long-range positional dependencies. Ultimately, this pipeline offers a reliable foundation for the automated annotation of eukaryotic proteomes, ensuring high-fidelity identification of the secretory machinery in an increasingly data-rich genomic landscape.

References

Almagro Armenteros, J.J. et al. (2019) SignalP 5.0 improves signal peptide predictions using deep neural networks. Nature Biotechnology, 37, 420–423.

Kyte, J. and Doolittle, R.F. (1982) A simple method for displaying the hydropathic character of a protein. Journal of Molecular Biology, 157, 105–132.

Nielsen, H. et al. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. Protein Engineering, Design and Selection, 10, 1–6.

Pedregosa, F. et al. (2011) Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830.

Steinegger, M. and Söding, J. (2017) MMseqs2 enables ultra-fast and sensitive sequence search and clustering. Nature Biotechnology, 35, 1026–1028.

The UniProt Consortium (2025) UniProt: the universal protein knowledgebase in 2025. Nucleic Acids Research, 53, D1.

von Heijne, G. (1983) Patterns of amino acids near signal-sequence cleavage sites. European Journal of Biochemistry, 133, 17–21.

von Heijne, G. (1986) A new method for predicting signal sequence cleavage sites. Nucleic Acids Research, 14, 4683–4690.