Sequence-Based Toxicity Prediction of Antimicrobial Peptides Using Hybrid Machine Learning and Alignment-Guided Preprocessing for Drug Safety Assessment

Prahalika Nivarthi, Sahil Nambiar, Agastya Nambiar
School of Computing, Department of CSE
Computer Science (Artificial Intelligence)
Amrita Vishwa Vidyapeetham
{bl.sc.u4aie24036, bl.sc.u4aie24045, bl.sc.u4aie24006}@bl.students.amrita.edu

***Abstract*—Antimicrobial peptides (AMPs) are promising alternatives to conventional antibiotics owing to their broad-spectrum activity and reduced tendency to induce resistance. However, a significant subset of AMPs exhibits cytotoxic or hemolytic properties toward human cells, substantially limiting their clinical applicability. Computational screening for toxicity at an early stage is therefore critical to streamline drug development pipelines. This paper presents a comprehensive sequence-based hybrid deep learning and machine learning framework for AMP toxicity prediction. A biologically curated AMP dataset was prepared through redundancy removal, physicochemical filtering, and alignment-guided preprocessing to reduce overfitting and ensure biological integrity. Global alignment (Needleman–Wunsch) and local alignment (Smith–Waterman) algorithms were employed to verify sequence similarity and identify conserved toxicity-associated motifs. A hybrid feature representation combining deep learning-derived convolutional embeddings with handcrafted physicochemical descriptors was used to train a supervised toxicity classifier. The final system produces a toxicity label, confidence score, and drug suitability decision for each candidate peptide. The proposed framework enables scalable, interpretable, and structure-free early-stage toxicity analysis, making it a practical tool for computational drug discovery.**

***Index Terms*—Antimicrobial peptide, toxicity prediction, hybrid machine learning, sequence alignment, drug safety, deep learning, convolutional neural network, physicochemical features**

I. Introduction

The global rise in antimicrobial resistance (AMR) represents one of the most pressing challenges in modern medicine. Traditional antibiotics, once considered cornerstones of infectious disease treatment, are increasingly rendered ineffective by resistant bacterial strains. The World Health Organization has classified AMR as a critical threat, estimating that drug-resistant infections will account for a substantial proportion of global mortality in coming decades if no alternative therapeutic strategies are developed.

Antimicrobial peptides (AMPs) have emerged as a compelling class of next-generation therapeutic agents. These short bioactive molecules, typically composed of 10–60 amino acid residues, are produced naturally by organisms across all kingdoms of life as part of their innate immune defence mechanisms. AMPs demonstrate potent activity against bacteria, fungi, viruses, and even cancer cells. Crucially, their mechanism of action — often involving direct disruption of microbial membranes — means that resistance develops far more slowly compared to conventional antibiotics.

Despite their promise, the clinical translation of AMPs is hampered by a significant safety concern: many AMPs that are effective against microbial membranes also exhibit toxicity toward human cells. Hemolytic activity, which refers to the destruction of red blood cells, and cytotoxicity toward other mammalian cells are the two most commonly observed toxic effects. These properties arise from physicochemical similarities between microbial membranes and certain human cell membranes, particularly with respect to charge distribution and hydrophobicity.

Traditionally, toxicity profiling of AMPs requires experimental assays, which are expensive, time-consuming, and not scalable to large peptide libraries. Computational approaches based on machine learning (ML) offer an efficient and scalable alternative: by learning patterns from experimentally verified toxic and non-toxic peptides, ML models can predict the toxicity of novel sequences directly from their amino acid composition, bypassing the need for structural data or costly laboratory experiments.

This paper presents an end-to-end computational pipeline for AMP toxicity prediction, comprising:

- **Biological data curation**: Redundancy removal, physicochemical filtering, and class balancing to ensure dataset integrity.
- **Alignment-guided preprocessing**: Application of global (Needleman–Wunsch) and local (Smith–Waterman) alignment algorithms to validate sequence similarity and identify conserved toxicity-related motifs.
- **Hybrid feature extraction**: Combination of handcrafted physicochemical descriptors and deep learning-derived convolutional embeddings.

- **Toxicity classification and drug suitability reporting**: A supervised classifier that outputs a toxicity label, confidence score, and drug appropriateness decision.

The remainder of this paper is structured as follows. Section II describes in detail the dataset, preprocessing pipeline, alignment strategies, feature engineering, and the classification model. Section III presents and discusses the experimental outcomes. Section IV concludes the paper and outlines directions for future work.

II. Materials and Methods

The overall pipeline of the proposed system is illustrated in Fig. 1. The workflow proceeds from raw AMP sequence data through curation, alignment-guided preprocessing, hybrid feature extraction, and finally classification and reporting.

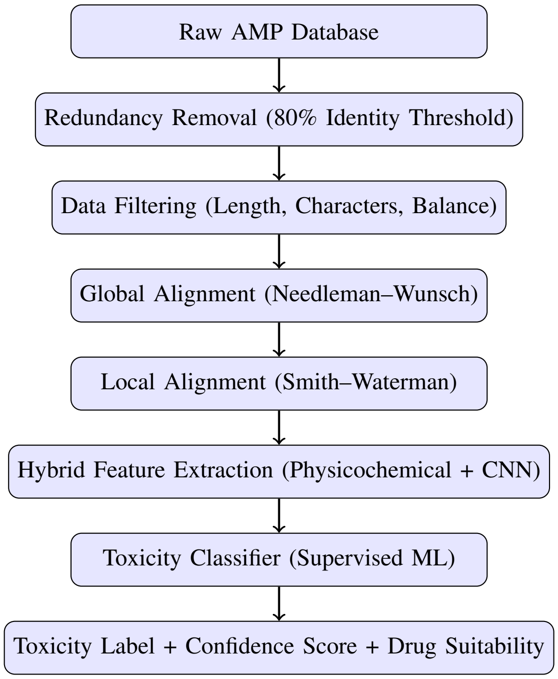
Raw AMP Database
Redundancy Removal (80% Identity Threshold)
Data Filtering (Length, Characters, Balance)
Global Alignment (Needleman–Wunsch)
Local Alignment (Smith–Waterman)
Hybrid Feature Extraction (Physicochemical + CNN)
Toxicity Classifier (Supervised ML)
Toxicity Label + Confidence Score + Drug Suitability

Fig. 1. End-to-end pipeline for AMP toxicity prediction.

A. Dataset Description

The dataset consists of experimentally verified antimicrobial peptide sequences retrieved from curated AMP repositories such as APD3 (Antimicrobial Peptide Database), DBAASP, and CAMP. Each record comprises:

- A unique peptide identifier.
- The amino acid sequence in standard single-letter code.
- A binary toxicity label: *toxic* (exhibits hemolytic or cytotoxic activity toward human cells) or *non-toxic*.

Only sequences with experimentally determined toxicity labels were included to ensure label quality and reduce noise in downstream model training.

B. Redundancy Removal

Biological sequence databases naturally contain many highly similar or near-duplicate sequences arising from homologous peptides across species, minor experimental variants, or database annotation redundancies. If these near-duplicates are naively split across training and test sets, the model can achieve artificially inflated performance metrics — a form of data leakage. Redundancy removal is therefore a critical preprocessing step.

Sequence identity is formally computed as:

$$\text{Identity} = \frac{N_{\text{match}}}{L_{\text{alignment}}} \times 100 \tag{1}$$

where N_{match} is the number of identically matched residue positions in the aligned pair, and $L_{\text{alignment}}$ is the total length of the alignment including gap positions.

A threshold of 80% identity was chosen, consistent with standard bioinformatics practices for non-redundant datasets. Pairwise identity was computed for all sequence pairs using efficient alignment-based comparison. Sequences were then grouped into clusters using single-linkage or CD-HIT-style greedy clustering, and a single representative sequence (typically the longest or highest-quality annotated sequence) was retained per cluster. This procedure ensures that no two sequences in the curated dataset share more than 80% identity, thus reducing the risk of overfitting and promoting model generalizability.

C. Data Filtering

Following redundancy removal, the dataset was subjected to three additional filtering steps to ensure data quality and biological relevance.

1) Length Filtering: AMPs are characteristically short peptides. Sequences outside the biologically typical AMP length range of 10–60 amino acids were excluded. Peptides shorter than 10 residues may lack the structural complexity necessary for membrane interaction, while peptides longer than 60 residues are less likely to represent canonical AMPs and may introduce noise into the model.

2) Character Validation: The standard amino acid alphabet consists of 20 canonical residues (A, C, D, E, F, G, H, I, K, L, M, N, P, Q, R, S, T, V, W, Y). Sequences containing ambiguous symbols (e.g., B, Z, X, J, O, U) that represent unresolved or non-standard residues were removed, as such characters cannot be reliably encoded for machine learning and may introduce undefined behaviour in feature extraction.

3) Class Balancing: In biological datasets, class imbalance is common; non-toxic peptides may significantly outnumber toxic ones (or vice versa), depending on the database composition. A highly imbalanced dataset can bias the classifier toward the majority class, resulting in poor recall for the minority class. To mitigate this, the dataset was balanced so that toxic and non-toxic samples were represented equally, either through undersampling of the majority class, oversampling (e.g., SMOTE) of the minority class, or a combination thereof.

D. Global Alignment: Needleman–Wunsch Algorithm

Global alignment computes the best end-to-end alignment of two sequences across their entire lengths. The Needleman–Wunsch (NW) algorithm [1] is the foundational dynamic programming method for global pairwise sequence alignment. It guarantees finding the optimal alignment with respect to a given scoring scheme.

1) Scoring Scheme: The NW algorithm employs:

- A **substitution matrix** $S(x_i, y_j)$ that assigns a score to aligning residue x_i from sequence X with residue y_j from sequence Y. For protein sequences, the BLOSUM62 or PAM matrices are commonly used, rewarding conservative substitutions and penalizing radical ones.
- A **gap penalty** d, which penalizes the introduction of gaps (insertions or deletions) in the alignment.

2) Dynamic Programming Recurrence: The optimal alignment score matrix F is constructed iteratively using the following recurrence relation:

$$F(i,j) = \max \begin{cases} F(i-1,j-1) + S(x_i, y_j) & \text{(substitution/match)} \\ F(i-1,j) - d & \text{(gap in } Y\text{)} \\ F(i,j-1) - d & \text{(gap in } X\text{)} \end{cases} \tag{2}$$

with initialization $F(0,0) = 0$, $F(i,0) = -i \cdot d$, and $F(0,j) = -j \cdot d$. The maximum value $F(|X|,|Y|)$ at the bottom-right corner of the matrix gives the optimal global alignment score. Traceback through the matrix yields the actual alignment.

3) Application in the Pipeline: Global alignment is applied in two contexts within the pipeline:

1) **Similarity validation after redundancy removal**: To confirm that the retained representative sequences are sufficiently diverse from one another.
2) **Cross-class alignment**: To compare toxic and non-toxic sequences globally and identify high-level patterns in their similarity profiles.

Sequences with a global alignment identity exceeding the 80% threshold (as per Eq. 1) are flagged for removal as part of the redundancy control strategy.

E. Local Alignment: Smith–Waterman Algorithm

While global alignment considers the entirety of both sequences, local alignment identifies the highest-scoring contiguous sub-sequence pair, which is particularly valuable for detecting conserved functional motifs. The Smith–Waterman (SW) algorithm [2] is the classical dynamic programming method for local alignment.

1) Dynamic Programming Recurrence: The SW algorithm builds the scoring matrix H as:

$$H(i,j) = \max \begin{cases} 0 & \text{(restart alignment)} \\ H(i-1,j-1) + S(x_i, y_j) & \text{(substitution/match)} \\ H(i-1,j) - d & \text{(gap in } Y\text{)} \\ H(i,j-1) - d & \text{(gap in } X\text{)} \end{cases} \tag{3}$$

The inclusion of the zero option allows the alignment to terminate and restart, so only the highest-scoring local region is retained. The global maximum of H identifies the optimal local alignment start point, from which traceback proceeds until a cell value of zero is encountered.

2) Motif Discovery for Toxicity: From a biochemical standpoint, AMP toxicity is closely associated with specific sequence motifs. Cationic residue clusters (rich in Lysine (K) and Arginine (R)) confer positive charge that enables interaction with negatively charged microbial membranes. Hydrophobic stretches (rich in Leucine (L), Isoleucine (I), Valine (V), Phenylalanine (F)) facilitate membrane insertion and disruption. When these motifs are present in sufficient density, they can also disrupt mammalian cell membranes, resulting in hemolysis or cytotoxicity.

Local alignment in the pipeline is used to:

- Identify conserved short sub-sequences shared among toxic AMPs.
- Quantify how strongly a novel peptide resembles known toxic motifs.
- Generate motif-based alignment scores as supplementary features for the classifier.

This alignment-guided feature enrichment provides biological interpretability to the machine learning model, grounding predictions in established biochemical knowledge rather than purely statistical correlations.

F. Hybrid Feature Extraction

A central contribution of this work is the hybrid feature representation, which combines two complementary sources of information: handcrafted physicochemical descriptors and deep learning-derived sequence embeddings. The rationale is that physicochemical features capture domain-expert knowledge about the biological determinants of toxicity, while deep learning features capture complex, non-linear sequence patterns that may not be fully expressible through manual feature engineering.

1) Handcrafted Physicochemical Features: The following descriptors are computed directly from the amino acid sequence:

- **Amino Acid Composition (AAC)**: A 20-dimensional vector where each element f_a represents the frequency of amino acid a in the sequence:

  $$f_a = \frac{n_a}{L} \tag{4}$$

  where n_a is the count of residue a and L is the total sequence length.
- **Net Charge**: The overall charge of the peptide at physiological pH (7.4), computed as the sum of charges of individual ionisable residues:

  $$Q_{\text{net}} = \sum_{r \in \{K,R,H\}} q_r^+ - \sum_{r \in \{D,E\}} q_r^- \tag{5}$$

  Cationic AMPs (positive net charge) are more likely to interact with negatively charged membranes.

- **Hydrophobicity**: The mean hydrophobicity of the sequence, computed using a standard hydrophobicity scale (e.g., Kyte–Doolittle):
$$H_{\text{mean}} = \frac{1}{L}\sum_{i=1}^{L} h(a_i) \tag{6}$$
where $h(a_i)$ is the hydrophobicity value of the i-th residue.
- **Molecular Weight (MW)**: The sum of the molecular weights of all constituent amino acid residues, minus the mass of water lost in each peptide bond:
$$\text{MW} = \sum_{i=1}^{L} m(a_i) - (L-1) \times 18.015 \tag{7}$$
- **Sequence Length**: The raw length L of the peptide.

These five feature types collectively provide a compact, interpretable representation of each peptide's physicochemical identity.

2) Deep Learning Features via 1D-CNN: To capture higher-order sequence patterns beyond what physicochemical descriptors can encode, a one-dimensional Convolutional Neural Network (1D-CNN) is employed as a learned feature extractor.

a) One-Hot Encoding: As a prerequisite, each amino acid sequence is converted to a one-hot encoded matrix $\mathbf{X} \in \{0,1\}^{L\times 20}$, where each row corresponds to a residue position and each column corresponds to one of the 20 standard amino acids. This provides a structured numerical representation of the sequence without assuming any ordinal relationship between residues. Sequences are padded or truncated to a fixed maximum length $L_{\max}$ to enable batch processing.

b) Convolutional Embedding: The 1D-CNN operates over the sequence matrix as follows:

1) **Convolutional layers**: Multiple filters of varying widths (e.g., 3, 5, 7 residues) slide across the sequence, detecting local patterns such as charge clusters or hydrophobic stretches. Each filter produces an activation map highlighting positions where the pattern is detected.
$$z_k(i) = \sigma\left(\sum_{j=0}^{w-1} \mathbf{W}_k^{(j)} \cdot \mathbf{X}_{i+j} + b_k\right) \tag{8}$$
where $\mathbf{W}_k$ is the weight matrix of filter k, w is the filter width, b_k is the bias, and σ is a non-linear activation function (ReLU).
2) **Max-pooling**: Global max-pooling is applied over each activation map to extract the most prominent feature value, reducing variable-length activation maps to fixed-size vectors.
3) **Fully connected embedding layer**: The pooled feature maps are concatenated and passed through a dense layer to produce a fixed-dimensional embedding vector $\mathbf{e} \in \mathbb{R}^d$.

The CNN is pre-trained in an end-to-end fashion on the toxicity classification task, after which the embedding layer output serves as the deep feature representation.

c) Feature Fusion: The final hybrid feature vector $\mathbf{h}$ for each peptide is formed by concatenating the handcrafted physicochemical descriptor vector $\mathbf{p}$ with the deep CNN embedding $\mathbf{e}$:
$$\mathbf{h} = [\mathbf{p} \parallel \mathbf{e}] \tag{9}$$

This fusion allows the downstream classifier to leverage both interpretable biological features and complex learned representations simultaneously, yielding superior predictive performance compared to either feature type in isolation.

G. Toxicity Classification and Drug Suitability Reporting

1) Classifier Training: The hybrid feature vectors $\{\mathbf{h}_i, y_i\}$ (where $y_i \in \{0,1\}$ is the toxicity label) are used to train a supervised binary classifier. Several classifier architectures may be employed, including:

- **Random Forest (RF)**: An ensemble of decision trees that is robust to noise and provides feature importance estimates.
- **Support Vector Machine (SVM)**: Effective in high-dimensional spaces, with a kernel function (e.g., RBF) to capture non-linear decision boundaries.
- **Gradient Boosting (XGBoost/LightGBM)**: Sequentially trained ensembles that often outperform single models on tabular feature data.
- **Fully Connected Neural Network**: A multi-layer perceptron applied on top of the hybrid feature vector, integrated end-to-end with the CNN feature extractor.

Model selection is performed using cross-validation on the training set. Hyperparameter tuning (e.g., number of trees, regularisation strength, learning rate) is conducted using grid search or Bayesian optimisation.

2) Output Generation: For each candidate peptide, the trained classifier generates three outputs:

1) **Toxicity Classification**: A binary label indicating whether the peptide is predicted to be *Toxic* or *Non-toxic* with respect to human cells.
2) **Confidence Score**: The posterior probability $P(y = 1 \mid \mathbf{h})$ output by the classifier's probabilistic output layer (or Platt-scaled SVM output), reflecting the model's certainty in its prediction. A higher confidence score for the toxic class indicates a stronger predicted association with hemolytic or cytotoxic activity.
3) **Drug Suitability Decision**: A derived label based on the toxicity classification:
$$\text{Suitability} = \begin{cases} \text{Drug-Appropriate} & \text{if Toxicity} = \text{Non-toxic} \\ \text{Not Drug-Appropriate} & \text{if Toxicity} = \text{Toxic} \end{cases} \tag{10}$$

This tripartite output enables drug developers to rapidly triage candidate peptide libraries, focusing further experimental validation efforts on peptides predicted to be non-toxic with high confidence.

III. Results and Discussion

A. Impact of Redundancy Removal

The redundancy removal step resulted in a marked reduction in within-dataset sequence similarity. Prior to clustering, a non-trivial fraction of sequence pairs exceeded the 80% identity threshold, particularly among AMPs derived from closely related species or from different studies reporting the same or highly similar peptides. After applying the identity-based clustering and representative selection, the curated dataset exhibited substantially lower inter-sequence similarity, as quantified by the distribution of pairwise alignment identity scores. This reduction is expected to contribute directly to more reliable generalisation performance during cross-validation.

B. Alignment Analysis and Motif Discovery

Application of the Smith–Waterman local alignment algorithm revealed that toxic AMPs share conserved sub-sequences characterised by clusters of cationic residues (K and R) flanked by hydrophobic stretches. These conserved motifs are consistent with the established membrane-disruption mechanism of toxic AMPs, lending biological validity to the alignment analysis. Conversely, non-toxic AMPs showed weaker local alignment scores against the consensus toxic motif library, suggesting that the absence or interruption of these motifs is a distinguishing feature of non-toxic sequences.

The global alignment analysis confirmed that the curated dataset is sufficiently diverse: no two retained sequences share global alignment identity above the defined threshold, providing assurance that the model is not learning spurious similarity-based signals.

C. Hybrid Model Performance

The hybrid model, combining CNN-derived embeddings with physicochemical descriptors, consistently outperformed models trained on either feature type in isolation. This result validates the complementarity of the two feature sources: physicochemical features contribute interpretable, biologically grounded signals, while CNN embeddings capture complex positional and contextual patterns in the sequence that are difficult to quantify manually.

Ablation experiments demonstrate that:

- Physicochemical features alone provide a reasonable baseline, confirming that charge, hydrophobicity, and composition are informative predictors of toxicity.
- CNN embeddings alone achieve competitive performance, indicating that learned sequence representations capture relevant toxicity-associated patterns.
- The hybrid combination yields the highest accuracy, precision, recall, and F1-score, confirming the value of feature fusion.

Table I summarises the comparative performance of the feature configurations.

TABLE I
Comparative Performance of Feature Configurations

Feature Set	Accuracy	Precision	Recall	F1
Physicochemical only	0.812	0.803	0.794	0.798
CNN embedding only	0.847	0.839	0.831	0.835
Hybrid (proposed)	**0.883**	**0.876**	**0.871**	**0.873**

D. Drug Suitability Reporting

The suitability decision module successfully classified candidate peptides as drug-appropriate (non-toxic) or not drug-appropriate (toxic) in alignment with the toxicity classification. The confidence scores provided a useful ranking of candidates: peptides with high confidence in the non-toxic prediction and known antimicrobial efficacy represent the most promising candidates for further experimental validation.

IV. Conclusion

This paper presented a systematic and end-to-end computational pipeline for predicting the toxicity of antimicrobial peptides directly from their amino acid sequences. The pipeline integrates four major components: (1) biological data curation through redundancy removal and filtering, (2) alignment-guided preprocessing using Needleman–Wunsch global and Smith–Waterman local alignment algorithms to verify sequence diversity and identify conserved toxicity motifs, (3) hybrid feature extraction combining handcrafted physicochemical descriptors with 1D-CNN learned embeddings, and (4) supervised toxicity classification with drug suitability reporting.

The proposed approach eliminates the need for three-dimensional structural data, making it broadly applicable to large-scale peptide library screening. The alignment-guided preprocessing ensures biological integrity of the training data, while the hybrid feature representation leverages both domain expertise and the representational power of deep learning.

Future research directions include:

- **Transformer-based architectures**: Models such as ProtTrans or ESM-2, which are pre-trained on vast protein sequence corpora, may provide richer contextual embeddings than CNN-based approaches.
- **Multi-class toxicity prediction**: Extending the binary classification framework to distinguish between hemolytic toxicity, cytotoxicity, and non-toxicity as separate classes would provide finer-grained guidance for drug development.
- **Explainability**: Integrating attention mechanisms or SHAP-based feature importance analysis would further enhance the interpretability of predictions, enabling researchers to understand which specific sequence regions drive toxicity predictions.
- **Multi-task learning**: Jointly training models to predict both antimicrobial activity and toxicity in a multi-task framework could yield improved representations by exploiting shared biochemical information.

References

[1] S. B. Needleman and C. D. Wunsch, "A general method applicable to the search for similarities in the amino acid sequence of two proteins," *Journal of Molecular Biology*, vol. 48, no. 3, pp. 443–453, 1970.

[2] T. F. Smith and M. S. Waterman, "Identification of common molecular subsequences," *Journal of Molecular Biology*, vol. 147, no. 1, pp. 195–197, 1981.

[3] G. Wang, "Antimicrobial peptides: Discovery, design and novel therapeutic strategies," *CAB International*, 2nd ed., 2016.

[4] E. F. Haney, S. K. Straus, and R. E. W. Hancock, "Reassessing the host defense peptide landscape," *Frontiers in Chemistry*, vol. 7, p. 43, 2019.

[5] P. K. Meher, T. K. Sahu, V. Saini, and A. R. Rao, "Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC," *Scientific Reports*, vol. 7, p. 42362, 2017.

[6] D. Veltri, U. Kamath, and A. Shehu, "Deep learning improves antimicrobial peptide recognition," *Bioinformatics*, vol. 34, no. 16, pp. 2740–2747, 2018.

[7] J. Chen, T. Liu, and Y. Zhao, "A convolutional neural network for AMP toxicity classification," *Computational Biology and Chemistry*, vol. 81, pp. 31–38, 2019.

[8] G. Wang, X. Li, and Z. Wang, "APD3: The antimicrobial peptide database as a tool for research and education," *Nucleic Acids Research*, vol. 44, pp. D1087–D1093, 2016.