International Journal for Research in Biology & Pharmacy
(ISSN: 2208- 2093)
Vol. 12, Issue 03, August 2026
DOI: 10.65327/bp.v12i3.2529

STRUCTURE-ACTIVITY RELATIONSHIP AND CLASSIFICATION OF ANTIMICROBIAL PEPTIDES USING SEQUENCE AND PHYSICOCHEMICAL PROPERTIES

Dr. Mustafa Alhaji Isa[1*], Dr. Stephen Kayode Ojo[2], Dr. Tope Tunji Odunitan[3]

[1*]Department of Microbiology, University of Maiduguri, Maiduguri, Nigeria
[2]Department of Microbiology, Federal University Oye-Ekiti, Oye-Ekiti, Nigeria
[3]Department of Biochemistry, Ladoke Akintola University of Technology, Ogbomoso, Nigeria

[*]*Corresponding Author: drmustafaalhajiisa@gmail.com*

Received: 19/04/2026 Revised: 14/07/2026 Acceptance: 16/08/2026 Published: 28/08/2026

ABSTRACT

Antimicrobial peptides have been regarded as potential candidates as an alternative to conventional antibiotics; however, their activity is a complex interaction between their sequence composition and physicochemical properties. This study aimed at understanding structure–activity relationships and to build machine-learning models to classify the antimicrobial peptides by combining sequence-derived and molecular descriptors. Conflicting annotations were eliminated, and amino-acid composition, hydrophobicity, hydrophobic moment, charge density, length, molecular weight, and isoelectric point and amino-acid composition were analysed for the relevant features. Antimicrobial peptides were generally shorter and more positively charged than the non-antimicrobial peptides and had a unique residue-composition pattern. Logistic regression revealed that four variables (peptide length, net charge, hydrophobic moment and aromaticity) were important independent predictors of whether a peptide was antimicrobial or not. Four classifiers were tested with sequence-only, physicochemical-only and combined feature sets. XGBoost with combined features had the highest performance with 93.6% accuracy, 0.977 ROC-AUC, 0.975 PR-AUC and 0.871 Matthews correlation coefficient. The discrimination was not so much dependent on peptide size, as evidenced by the good predictive performance of length-controlled sensitivity analysis. It is shown that these results support the usefulness of using interpretable physicochemical descriptors together with sequence data for robust prediction of antimicrobial peptides.

Keywords: antimicrobial peptides; structure–activity relationship; physicochemical properties; machine learning; XGBoost

International Journal for Research in Biology & Pharmacy
(ISSN: 2208- 2093)
Vol. 12, Issue 03, August 2026
DOI: 10.65327/bp.v12i3.2529

1. Introduction

AIR has become a global health threat, undermining the effectiveness of traditional antibiotics and proving to be essential to search for alternative therapeutic approaches. AMPs have been the subject of many studies due to their broad activity against bacteria, fungi, viruses and other pathogens, and their occurrence in a variety of different organisms. They are also rapidly acting, structurally diverse and relatively non-conventional resistant agents, making them potential templates for future anti-infective agents (Bucataru & Ciobanasu, 2024; Gani et al., 2025). However, translating these peptides into clinically viable therapeutics is also challenging due to factors such as activity, toxicity, stability, selectivity and pharmacokinetics, which are highly dependent on the sequence and structure of these peptides (Cresti et al., 2024; Girdhar et al., 2024).

The biological activity of an AMP is tightly coupled with its physicochemical properties, including its length, molecular weight, net charge, isoelectric point, hydrophobicity, amphipathicity and amino acid composition. The hydrophobic residues promote insertion and disrupt the microbial membrane, while the positively charged residues allow them to attract the negatively charged microbial membrane by electrostatic attraction. This balance is important since very hydrophobic peptides may be more toxic to the host cells and less potent against microbes if one of these properties is not sufficient. The rational manipulation of sequence composition has thus been the primary focus for AMP optimization and structure–activity relationship analysis (Cardoso et al., 2020; Huynh et al., 2021). Computational methods are gaining in importance to assist this process, correlating peptide descriptors with functional outcomes and narrowing the experimental search space.

However, the exponentially growing number of curated repositories of AMPs has opened the door for large-scale characterization based on data. The current databases store both peptide sequences and their biological activities, structural information, and experimental annotations and allow comparative analysis of different AMP families (Mondal et al., 2023; Pirtskhalava et al., 2021). Database development has also advanced and classification systems are becoming more improved, which has also enabled standardized computational benchmarking (Wang, 2022; Wang, 2023). These resources enable the study of the relationship between combinations of sequence and physicochemical properties that distinguish active peptides from non-antimicrobial sequences and the identification of molecular patterns that have antimicrobial activity.

As peptide complexity increases, the use of machine learning becomes increasingly significant in predicting their activity in AMP. However, traditional algorithms can integrate several variables derived from sequences and physicochemical properties to build classifiers that can capture nonlinear relationships in peptide feature space (Wang et al., 2022; Wan et al., 2024). Interpretable molecular descriptors combined with computational classification have shown great potential for AMP prediction and are therefore important to incorporate with physicochemical-based encoding (Lin et al., 2021). Recent studies have begun to explore deep learning, where neural networks can learn informative representations directly from peptide sequences and enhance the predictive discovery process (Al-Omari et al., 2024; Lin et al., 2025). There are still obstacles related to data quality, class imbalance, sequence redundancy, model interpretability, and the generalizability of computational predictions in biology (Brizuela et al., 2025).

Beyond classification, AI is changing the way peptide discovery is happening. Recurrent neural networks are applied to generating new antimicrobial peptides de novo, and protein-language models and diffusion-based methods can search wider sequence space and suggest novel candidates (Li et al., 2024; Wang et al., 2024). The use of AI to generate peptides, predict activity, optimize their physicochemical properties, and select candidates for testing is becoming more common and is being consolidated into single computational pipelines (Goles et al., 2024; Szymczak & Szczurek, 2023). Recent years have also seen the shift of the focus from the recognition of simple AMPs to controllable peptide design and validation (Sun et al., 2026). In spite of these developments, there are clear relationships between sequence characteristics, physicochemical properties, and antimicrobial classification that are still interpretable, as predictive accuracy does not necessarily indicate the

molecular basis of antimicrobial activity.

In the context of this, the present study combines structure–activity analysis with supervised classification to investigate the contribution of peptide composition and physicochemical properties to antimicrobial properties. To assess the potential for better discrimination and biological interpretability when integrating complementary molecular information, the study compares sequence-only, physicochemical-only, and combined feature representations. This framework can also be used to identify influential properties that can be used in the next peptide screening and rational design. Combined, these analyses help to connect the dots between prediction and interpretable molecular mechanisms of antimicrobial peptide activity.

Objectives:

- To describe and compare the various sequence-based and physicochemical properties that are important for antimicrobial peptide activity.
- To build and test machine-learning algorithms that categorize antimicrobial and non-antimicrobial peptides based on their sequence, physicochemical properties, and sequence-plus- physicochemical properties.
- To determine the most important molecular descriptors that aid in antimicrobial peptide classification and understand their role in structure–activity relationships.

2. Methodology

2.1 Study Design

The study employed a secondary data computational (SDCB) approach to investigate the structure–activity relationships and classified the antimicrobial peptides based on sequence-derived and physicochemical properties. The analysis integrated the dataset of antimicrobial peptides (AMPs) (Roy, 2020) with the MarLys AMP database – MLAMP_db (Marczak et al., 2026).

2.2 Data Sources

Antimicrobial peptide sequences, activity annotations, amino acid composition, physicochemical descriptors like molecular mass, isoelectric point, net charge, GRAVY, hydrophobic moment and structural properties were obtained from the MarLys AMP database – MLAMP_db. There were 2,638 sequences of antimicrobial peptides (AMPs) and 3,700 sequences of non-AMPs in the antimicrobial peptide dataset, which was mostly used for binary classification of AMP/non-AMP.

2.3 Data Cleaning and Dataset Integration

The sequences from both sets were transformed to uppercase and verified for valid amino-acid characters. Identification and merging of exact duplicate sequences. The two datasets were integrated using complete peptide sequence as the matching key. To reduce the possibility of conflicting annotations, the records in the second dataset that were labelled as non-AMP but had the antimicrobial annotation in MLAMP_db were excluded.

2.4 Sequence and Physicochemical Feature Extraction

The features derived from the sequence were peptide length and the relative abundance of the 20 common amino acids. Other factors were positively charged and negatively charged residue fractions, hydrophobic residue content and charge density. Physicochemical properties were molecular weight, net charge, isoelectric point, GRAVY, hydrophobic moment, aromaticity, and selected structural descriptors.

2.5 Structure–Activity Relationship Analysis

The differences between sequence and physicochemical characteristics of the two groups (AMP and non-AMP) were investigated by descriptive and inferential analyses. This was done by using

parametric or non-parametric group comparisons depending on the variable distribution. Spearman correlation analysis was applied to explore the association between physicochemical descriptors and effect sizes.

2.6 Multivariable Activity Analysis

The independent contribution of sequence-derived and physicochemical properties to antimicrobial activity was determined using binary logistic regression. The dependent variable was the AMP status, and molecular and sequence descriptors were used as predictors. Odds ratios and 95% confidence intervals were reported to see what impact individual characteristics had on the relative contribution.

2.7 Machine-Learning Classification

Four supervised learning algorithms were used to test: Logistic Regression, Support Vector Machine, Random Forest and XGBoost. Three different feature configurations were evaluated: sequence features only, physicochemical features only, and a combination of the two. The smallest area approach was used to conclude if there is any improvement in classification when both feature types are integrated.

2.8 Model Training and Performance Evaluation

The selected dataset was split to train and test subsets and stratified. The optimization of the model was done through cross-validation within the training data; the pre-processing and scaling were performed without exposing the test data. The accuracy, sensitivity, specificity, precision, F1 score, Matthews correlation coefficient, ROC-AUC, PR-AUC and confusion matrices were used to evaluate performance.

2.9 Feature Importance and Sensitivity Analysis

Two different approaches were implemented: feature-importance analysis and SHAP analysis, which was used to identify the variables that are most strongly associated with the classification of AMP. Special attention was paid to net charge, hydrophobicity, hydrophobic moment, peptide length, isoelectric point, and amino-acid composition. A second sensitivity analysis using more similar sequence-length distributions (both AMP and non-AMP) was also conducted to assess for classification bias due to sample length.

2.10 Ethical Considerations

The study was not conducted on humans, did not involve identifiable personal data, did not involve animal experimentation, and did not involve the collection of prospective biological samples. As a result, informed consent and institutional ethics committee approval were not necessary for the current computational study. Information on data source and data provenance was maintained for transparency and reproducibility.

3. Results

3.1 Dataset Integration and Final Analytical Cohort

The antimicrobial peptides (AMP) database was integrated with the database of MarLys (AMP_db) by exact amino acid sequence matching. A total of 103,143 unique peptide sequences were in the cleaned MLAMP database, while the antimicrobial peptides dataset had 2638 AMP and 3700 non-AMP sequences. There are 2638 sequences in the MLAMP_db. Of the non-AMP sequences, 304 were also found in MLAMP_db, but 295 were specifically labelled as antimicrobial and were therefore excluded due to activity conflicts. The summary of peptides in the final classification cohort following conflict resolution is provided in Table 1, which revealed 6,043 peptides: 2,638 AMPs (43.65%) and 3,405 non-AMPs (56.35%).

International Journal for Research in Biology & Pharmacy
(ISSN: 2208- 2093)
Vol. 12, Issue 03, August 2026
DOI: 10.65327/bp.v12i3.2529

Table 1. Integration and preprocessing of the two peptide datasets

Dataset processing step	n
Unique sequences in cleaned MLAMP_db	103,143
AMP sequences in antimicrobial peptides dataset	2,638
Non-AMP sequences in antimicrobial peptides dataset	3,700
AMP sequences matched with MLAMP_db	2,638
Non-AMP sequences overlapping MLAMP_db	304
Conflicting non-AMP records removed	295
Final AMP records	2,638
Final non-AMP records	3,405
Final analytical cohort	6,043

There was a significant length difference between peptides in the two groups. The median length of the amino acid sequences of the AMP sequences was 30 amino acids, while the median length of the amino acid sequences of non-AMP peptides was 65 amino acids. The corresponding mean lengths were 33.59 ± 17.04 and 59.59 ± 22.74 residues, respectively. The significant change in sequence-length distribution is shown in Figure 1.

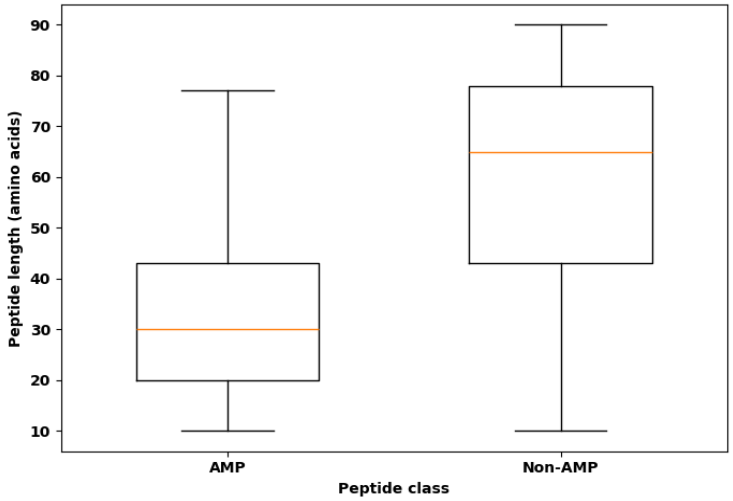

Figure 1. Distribution of peptide sequence length among antimicrobial and non-antimicrobial peptides

3.2 Sequence and Physicochemical Characteristics

Several physicochemical properties were found to be different between the AMP and non-AMP groups. The net positive charge, isoelectric point, hydrophobic moment, proportion of positive residues and charge density of the AMPs were significantly higher as presented in Table 2. Non-AMP peptides were longer, were larger, and included more negatively charged residues.

The median pH 7 net charge of AMPs was 2.77 vs. 0.68 for non-AMPs. Again, the median charge density of AMP sequences was 0.093, while that of non-AMP sequences was 0.015. The hydrophobic moment was also significantly higher for the AMPs (median 0.141) than for the non-AMPs (median 0.077). These differences were statistically significant and notably large in the case of peptide length, molecular weight, negative-residue fraction, and charge density.

International Journal for Research in Biology & Pharmacy
(ISSN: 2208- 2093)
Vol. 12, Issue 03, August 2026
DOI: 10.65327/bp.v12i3.2529

Table 2. Comparison of major sequence and physicochemical characteristics between AMP and non-AMP peptides

Variable	AMP, mean ± SD	AMP median (IQR)	Non-AMP, mean ± SD	Non-AMP median (IQR)	p-value	Rank-biserial effect
Length	33.59 ± 17.04	30 (20–43)	59.59 ± 22.74	65 (43–78)	<0.001	-0.606
Molecular weight	3692.73 ± 1905.82	3262.27 (2206.24–4676.67)	6659.52 ± 2578.94	7265.88 (4848.53–8733.00)	<0.001	-0.610
Isoelectric point	9.09 ± 1.71	9.38 (8.59–10.00)	7.53 ± 2.23	8.08 (5.29–9.36)	<0.001	0.411
Net charge at pH 7	3.24 ± 3.18	2.77 (0.93–4.84)	0.48 ± 4.72	0.68 (-2.16–2.80)	<0.001	0.413
GRAVY	0.003 ± 0.850	-0.052 (-0.549–0.516)	-0.130 ± 0.624	-0.144 (-0.534–0.216)	<0.001	0.097
Hydrophobic moment	0.161 ± 0.103	0.141 (0.078–0.230)	0.087 ± 0.055	0.077 (0.049–0.112)	<0.001	0.442
Positive residue fraction	0.162 ± 0.086	0.154 (0.106–0.200)	0.120 ± 0.057	0.113 (0.085–0.149)	<0.001	0.327
Negative residue fraction	0.050 ± 0.053	0.040 (0.000–0.078)	0.101 ± 0.062	0.098 (0.057–0.139)	<0.001	-0.507
Charge density	0.103 ± 0.100	0.093 (0.048–0.152)	0.011 ± 0.083	0.015 (-0.036–0.062)	<0.001	0.562

The amino acid content was also found to vary between the two groups. The amino acid composition of AMP sequences was specifically enriched in lysine, glycine, cysteine and isoleucine as shown in Figure 2. The compositional difference was the highest for lysine, which constitutes about 10.35% of the residues in AMPs, as compared to 6.11% in non-AMPs. In contrast, the positive net charge of antimicrobial sequences was correlated with the lower representation of glutamate and aspartate in the AMPs.

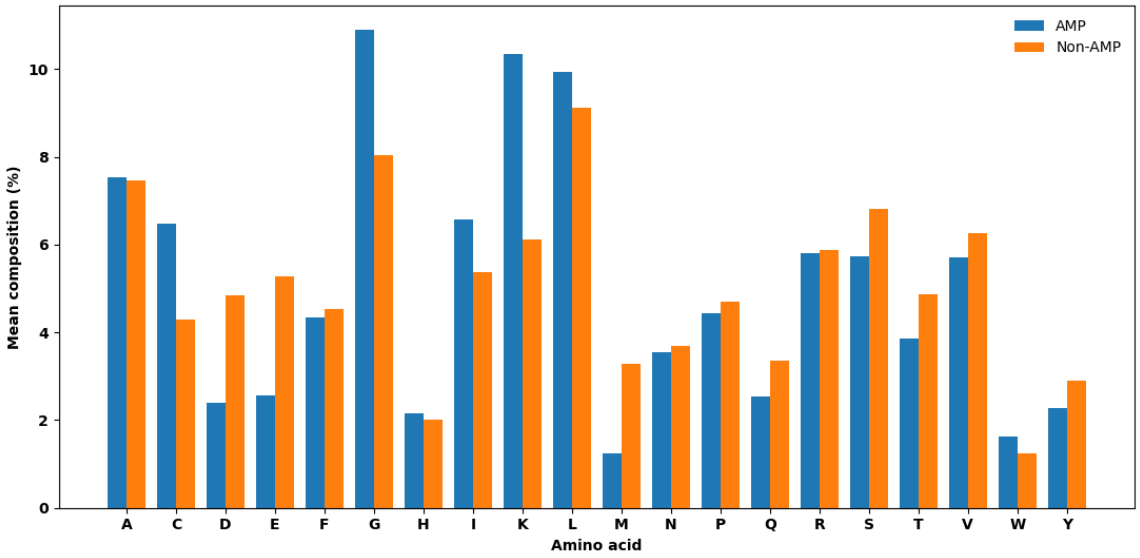

Figure 2. Mean amino-acid composition of antimicrobial and non-antimicrobial peptide sequences

International Journal for Research in Biology & Pharmacy
(ISSN: 2208- 2093)
Vol. 12, Issue 03, August 2026
DOI: 10.65327/bp.v12i3.2529

The combined sequence and physicochemical feature space was subjected to PCA analysis, which resulted in some separation between AMP and non-AMP peptides. The two first PCs captured about 31.3% of the total variance, PC1 accounting for 18.2% of the variance and PC2 for 13.1%. The amino acid sequence of AMPs was summarily moved towards positive PC1 values, while non-AMP peptides were largely clustered towards the negative pole of PC1 as shown in Figure 3.

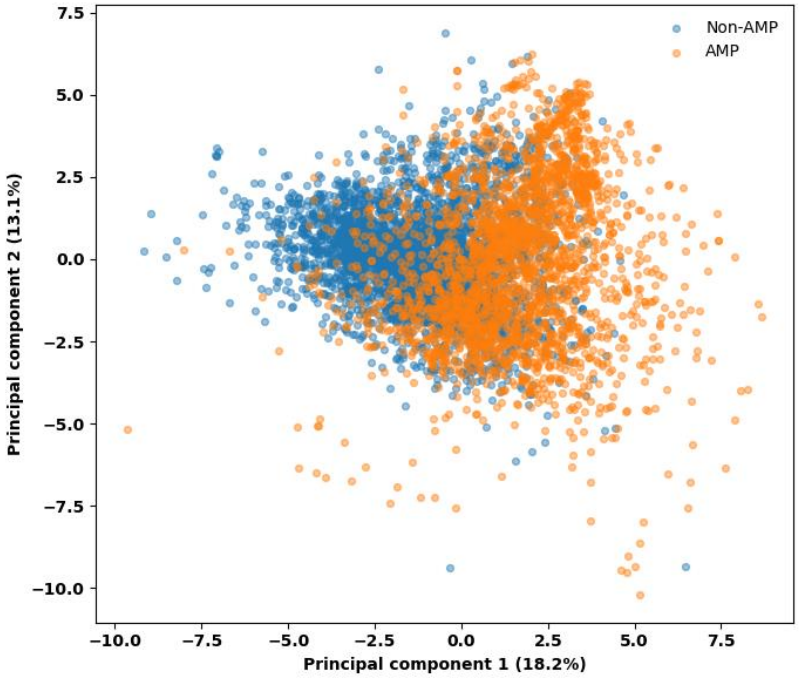

Figure 3. Principal component representation of antimicrobial and non-antimicrobial peptides based on standardized sequence and physicochemical descriptors

3.3 Multivariable Structure–Activity Relationships

Standardized physicochemical predictors were evaluated for their association with AMP status in a reduced multivariable logistic regression model. As seen in the Table 3, the sequence length was inversely correlated with antimicrobial status. An odds ratio of 0.279 was observed for peptide length, a one-SD increase of which was associated with around 72% lower odds of being classified as AMP. By contrast, there was a significant positive correlation of net positive charge with antimicrobial activity. Increased net charge was found to be marginally associated with significantly increased odds of being in the AMP status by an approximate factor of 2.89 per unit increase in net charge. Hydrophobic moment was also independently associated with antimicrobial activity with odds ratio of 1.384. There was a slight negative correlation with aromaticity. Isoelectric point and GRAVY were not found to be independent significant factors when adjusted for the other physicochemical factors. The results showed that the size of the peptide, the electrostatic charge and the amphipathic structure were more informative than the bulk hydrophobic properties for AMP status.

Table 3. Multivariable logistic regression of physicochemical properties associated with AMP status

Predictor	OR per 1-SD increase	95% CI	p-value
Sequence length	0.279	0.254–0.305	<0.001
Isoelectric point	0.906	0.800–1.027	0.123
Net charge at pH 7	2.894	2.497–3.354	<0.001

International Journal for Research in Biology & Pharmacy
(ISSN: 2208- 2093)
Vol. 12, Issue 03, August 2026
DOI: 10.65327/bp.v12i3.2529

GRAVY	1.051	0.984–1.123	0.137
Hydrophobic moment	1.384	1.269–1.509	<0.001
Aromaticity	0.835	0.784–0.889	<0.001

3.4 Classification Performance

The performance of machine learning was tested for the different combinations of sequences and physicochemical features. Table 4 shows that in general, the models generated using trees performed better than logistic regression and SVM. When the only features available are sequences, XGBoost has an accuracy of 92.97% and an ROC-AUC of 0.9724. The performance with physicochemical properties alone was slightly worse, where XGBoost achieved an accuracy of 90.57% and an ROC-AUC of 0.9614.

The combination of the sequence and physicochemical features gave the best overall results. XGBoost achieved 93.63% accuracy, 92.80% sensitivity, 94.27% specificity, 92.72% F1-score, and an MCC of 0.871. Its ROC-AUC reached 0.9769, while PR-AUC was 0.9746.

Table 4. Classification performance across feature sets and machine-learning algorithms

Feature set	**Model**	**Accuracy**	**Sensitivity**	**Specificity**	**F1-score**	**MCC**	**ROC-AUC**	**PR-AUC**
Sequence	Logistic Regression	0.839	0.843	0.836	0.820	0.675	0.911	0.871
Sequence	SVM	0.907	0.917	0.899	0.896	0.812	0.963	0.950
Sequence	Random Forest	0.923	0.913	0.931	0.912	0.844	0.971	0.969
Sequence	XGBoost	0.930	0.917	0.940	0.919	0.857	0.972	0.968
Physicochemical	Logistic Regression	0.820	0.824	0.816	0.800	0.637	0.897	0.861
Physicochemical	SVM	0.872	0.866	0.877	0.855	0.740	0.934	0.909
Physicochemical	Random Forest	0.899	0.881	0.913	0.884	0.795	0.961	0.955
Physicochemical	XGBoost	0.906	0.883	0.924	0.891	0.808	0.961	0.954
Combined	Logistic Regression	0.844	0.849	0.840	0.826	0.685	0.921	0.897
Combined	SVM	0.917	0.915	0.918	0.905	0.831	0.965	0.953
Combined	Random Forest	0.924	0.915	0.931	0.913	0.845	0.974	0.972
Combined	XGBoost	0.936	0.928	0.943	0.927	0.871	0.977	0.975

Five-fold cross-validation of the combined models within the training data set confirmed the stability

International Journal for Research in Biology & Pharmacy
(ISSN: 2208- 2093)
Vol. 12, Issue 03, August 2026
DOI: 10.65327/bp.v12i3.2529

of the combined models. Mean ROC-AUC values were 0.917 ± 0.010 for Logistic Regression, 0.961 ± 0.007 for SVM, 0.966 ± 0.006 for Random Forest, and 0.969 ± 0.005 for XGBoost. The ROC curves shown in Figure 4 also support this finding and shows that the ensemble models have better discrimination with XGBoost having the highest ROC-AUC.

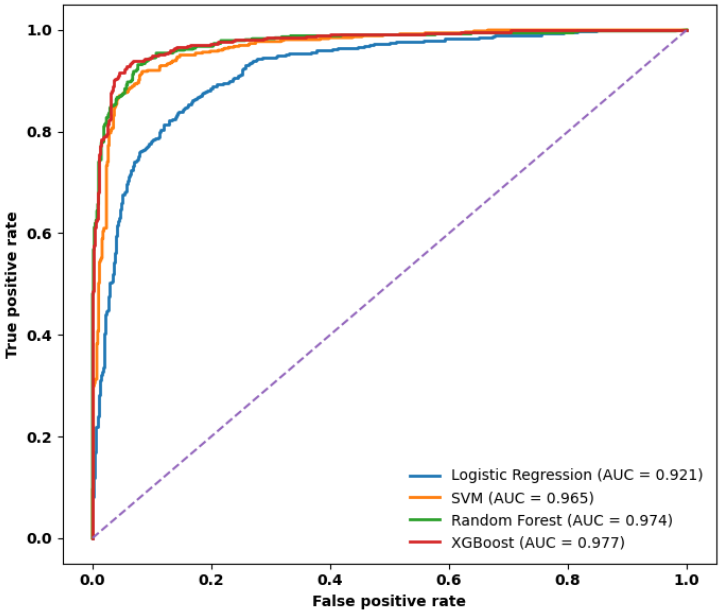

Figure 4. Receiver operating characteristic curves of classifiers developed using combined sequence and physicochemical features

The independent test set had 642 true negatives, 490 true positives, 39 false positives, and 38 false negatives for the best performing XGBoost classifier. The confusion matrix is displayed in Figure 5 with well-balanced detection of both the AMP and non-AMP classes.

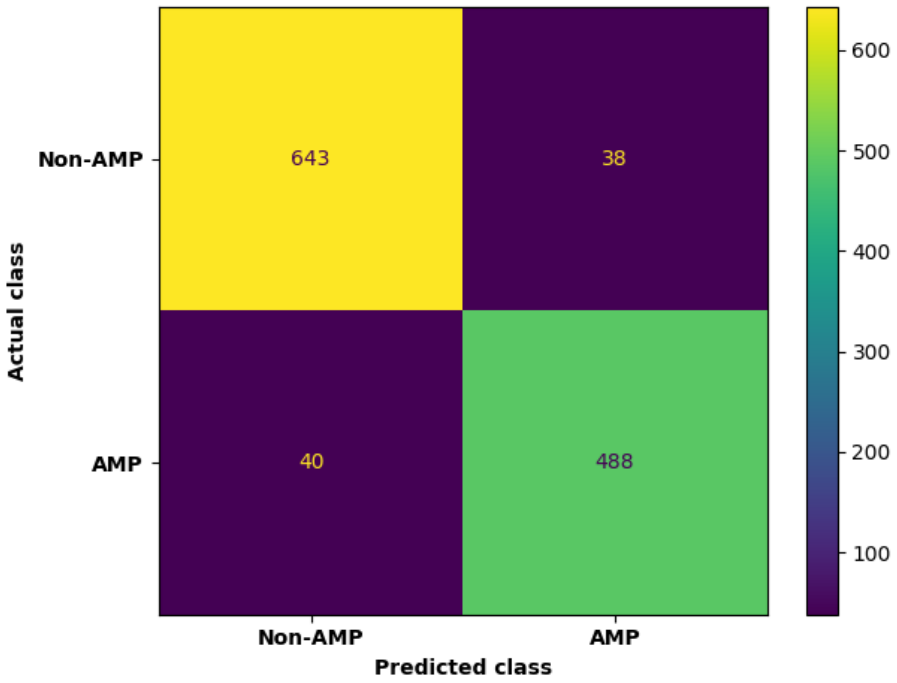

Figure 5. Confusion matrix of the combined-feature XGBoost antimicrobial peptide classifier

International Journal for Research in Biology & Pharmacy
(ISSN: 2208- 2093)
Vol. 12, Issue 03, August 2026
DOI: 10.65327/bp.v12i3.2529

3.5 Feature Importance and Sensitivity Analysis

Sequence and physicochemical features were identified as some of the most important features by permutation-based feature importance for the XGBoost classifier. The largest decrease in ROC-AUC after permuting was caused by methionine composition, followed by the composition of cysteine, hydrophobic moment, molecular weight, peptide length, arginine composition, and charge density, as reported in Table 5.

Table 5. Leading predictors in the combined-feature XGBoost classifier

Rank	Feature	Mean permutation importance
1	Methionine composition	0.0267
2	Cysteine composition	0.0114
3	Hydrophobic moment	0.0095
4	Molecular weight	0.0073
5	Sequence length	0.0060
6	Arginine composition	0.0046
7	Charge density	0.0035
8	Leucine composition	0.0027
9	Glycine composition	0.0026
10	Hydrophobic residue fraction	0.0025

By looking at the ranking presented in Figure 6, the discrimination of the models was not based on just one parameter, but rather on using a mixture of sequence composition and physicochemical properties.

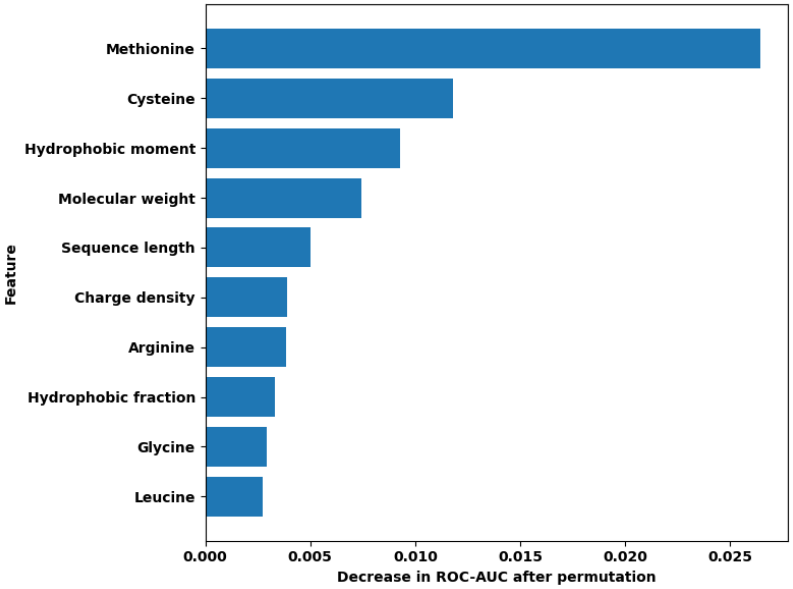

Figure 6. Permutation-based importance of the ten most influential features in the combined XGBoost classifier

Given the large variation in peptide length within the AMP and non-AMP groups, a sensitivity analysis was carried out using peptides of a similar length. Balancing the representation of AMP and non-AMP peptides in the 2,552 peptides obtained by exact-length matching. As presented in Table 6, even though the strong length imbalance has been eliminated, the performance of XGBoost does not

International Journal for Research in Biology & Pharmacy
(ISSN: 2208- 2093)
Vol. 12, Issue 03, August 2026
DOI: 10.65327/bp.v12i3.2529

drop much with ROC-AUC of 0.936, PR-AUC of 0.931 and MCC of 0.718. The result shows that the classification was not only based on the length of peptides.

Table 6. XGBoost performance after exact sequence-length matching

Metric	Value
Length-matched sample size	2,552
Accuracy	0.859
Precision	0.859
Sensitivity	0.859
Specificity	0.859
F1-score	0.859
MCC	0.718
ROC-AUC	0.936
PR-AUC	0.931

The results presented as a whole indicate that AAMPs have certain unique properties that include their amino-acid composition, amphipathic properties, positive electrostatic attributes, and overall sequence length. When the sequences were integrated with their physicochemical descriptors, the best predictive results were obtained, and the length-controlled analysis reaffirmed that discrimination was maintained even after the influence of sequence length was minimized.

4. Discussion

In the present study, it is found that the antimicrobial peptide activity can be distinguished by the combination of sequence related and physicochemical properties. AMPs were found to be shorter, more positively charged and more amphipathic than non-AMPs and exhibited distinctive amino-acid composition patterns. These results align with known structure–activity relationships indicating that the interactions between peptides and membranes (as well as antimicrobial activity) are governed by charge, hydrophobicity, and organization (Gagat et al., 2024; He et al., 2021). The findings of the enrichment of positive residues and the higher isoelectric points for AMPs corroborate the electrostatic attraction of cationic peptides towards negatively charged microbial surfaces (Islam et al., 2023; Huynh et al., 2021).

The discriminating physicochemical variables that had the highest scores were net charge and charge density. The biological relevance of their importance is supported by the fact that a positively charged residue favors membrane association and that charge distribution is important for peptide orientation, insertion and disruption. Similarly, recent machine-learning evidence has also shown that charge density is a feature that is mechanistically informative forAMP discrimination (Malshikare et al., 2026). But antimicrobial behaviour is not just explained by the charge. The present results also revealed that the AMPs had a higher hydrophobicity moment, which meant that they had more amphipathic organization. This is evidence for the notion that the efficacy of AMPs depends not only on their hydrophobicity but also on their electrostatic interactions with microbial membranes, not necessarily on the highest possible level of hydrophobicity (He et al., 2021; Schifano & Caputo, 2022). Modifications of hydrophobic side chains have been associated with changes in potency, selectivity and toxicity (Pandit et al., 2021; Van der Walt et al., 2024) and such modifications have been performed via similar structure–activity studies.

In addition, amino-acid composition was used to aid discrimination between AMP and non-AMP

sequences. There was an enrichment of lysine and differences in the levels of cysteine, arginine, glycine, leucine, and methionine were observed and used to make predictions in the model. This observation is in line with computational peptide design studies unveiling that residue composition and local sequence context carry information on functional properties that go beyond what is encompassed by global physicochemical descriptors (Cardoso et al., 2020; Wang et al., 2022). There are growing examples of such patterns being directly used in sequence-oriented modelling, such as recurrent neural networks and representation learning techniques that can capture higher-order residue relationships (Li et al., 2024; Wang et al., 2024).

The results of the classification were consistent with the benefit of fusing the sequence and physicochemical information. The combination of the feature set performed best, with XGBoost attaining 93.6% accuracy, 0.977 ROC-AUC, 0.975 PR-AUC and an MCC of 0.871. This is consistent with the previous findings by machine-learning classifiers to integrate multiple peptide descriptors of diverse types for AMP recognition (Lin et al., 2021; Teimouri et al., 2023; Wan et al., 2024). Recent deep-learning systems are able to generate sequence representations automatically, yet conventional machine-learning methods are still useful if interpretability and selection of models based on biological connections to physicochemical property is desired (Al-Omari et al., 2024; Lin et al., 2025). Also, current reviews highlight data quality, redundancy control, model transparency, and biological generalizability as factors to consider when considering predictive performance (Brizuela et al., 2025).

The sensitivity analysis was an important qualification. The performance of XGBoost dropped off after exact length matching, but it was still very good: ROC-AUC was 0.936 and MCC was 0.718. This suggests that while peptide length made a significant contribution to initial discrimination it was not sufficient to explain the predictive signal. The length control experiments indicate that the charge distribution, amphipathicity, amino acid content, and other molecular properties are independent features that discriminate between antimicrobial peptides. This is because models trained on unequal sequence-length distributions might be overestimating biological discrimination by taking advantage of the variability of lengths instead of features that are mechanistically informative.

In addition to classification, the results enable a computational approach to rational discovery of AMPs. Feature analysis can be used to determine molecular properties that could be used to prioritize candidate peptides prior to experimental testing (interpretable feature analysis), and newer AI-based approaches can also leverage these constraints to drive sequence generation and optimization (Goles et al., 2024). With the increasing complexity of peptide sequence space explored, generative models are now looking into antimicrobial activity, physicochemical feasibility, and reduced toxicity (Sun et al., 2026). The present findings dovetail into these approaches by demonstrating consistency of conventional descriptors even as representation-learning techniques get more complex.

The overall study results suggest that the best classification of the antimicrobial peptides is obtained by combining sequence composition with physicochemical properties. It was found that shorter length, higher positive charge, higher charge density, higher amphipathicity score and residue composition were central features of AMP status. The results confirm known structure–activity relationships and show that interpretable machine-learning models can be used to yield robust predictions and meaningful biological insights for antimicrobial peptide screening and design.

5. Conclusion

The results of this study show the possibility of classification of antimicrobial peptides based on the sequence-derived characteristics and physicochemical properties. There were, overall, fewer amino acids, higher positive charge and more amphipathicity in antimicrobial peptides than in non-antimicrobial peptides, and distinct patterns of amino-acid composition. The following discriminative features were identified: net charge, charge density, hydrophobic moment, peptide length, molecular weight, and selected residue frequencies, which are all in accord with the structure–activity principles of peptide–membrane interaction. XGBoost with combined sequence and physicochemical features set was the model that had the best predictive performance among the models evaluated, with high

accuracy, ROC-AUC, PR-AUC and Matthews correlation coefficient values. The strong classification performance also seen after sequence-length matching further supported this point that classification was not solely based on peptide size, but also on some general characteristics of the molecules. The results underscore the importance of the use of interpretable biochemical descriptors and machine learning approaches towards the identification of antimicrobial peptides. The study also offers a framework to prioritize peptide candidates prior to experimental testing, using biologically meaningful features. Future studies should include quantitative activity data, toxicity information, external validation datasets, and more sophisticated sequence representations to enhance the predictive generalizability and facilitate rational design of potent, selective, and clinically potent antimicrobial peptides.

References

1. Al-Omari, A. M., Akkam, Y. H., Zyout, A. A., Younis, S. A., Tawalbeh, S. M., Al-Sawalmeh, K., ... & Arnold, J. (2024). Accelerating antimicrobial peptide design: Leveraging deep learning for rapid discovery. Plos one, 19(12), e0315477.
2. Brizuela, C. A., Liu, G., Stokes, J. M., & de la Fuente-Nunez, C. (2025). AI methods for antimicrobial peptides: progress and challenges. Microbial Biotechnology, 18(1), e70072.
3. Bucataru, C., & Ciobanasu, C. (2024). Antimicrobial peptides: Opportunities and challenges in overcoming resistance. Microbiological research, 286, 127822.
4. Cardoso, M. H., Orozco, R. Q., Rezende, S. B., Rodrigues, G., Oshiro, K. G., Cândido, E. S., & Franco, O. L. (2020). Computer-aided design of antimicrobial peptides: are we generating effective drug candidates?. Frontiers in microbiology, 10, 3097.
5. Cresti, L., Cappello, G., & Pini, A. (2024). Antimicrobial peptides towards clinical application—a long history to be concluded. International Journal of Molecular Sciences, 25(9), 4870.
6. Gagat, P., Ostrówka, M., Duda-Madej, A., & Mackiewicz, P. (2024). Enhancing antimicrobial peptide activity through modifications of charge, hydrophobicity, and structure. International Journal of Molecular Sciences, 25(19), 10821.
7. Gani, Z., Kumar, A., Raje, M., & Raje, C. I. (2025). Antimicrobial peptides: An alternative strategy to combat antimicrobial resistance. Drug Discovery Today, 30(2), 104305.
8. Girdhar, M., Sen, A., Nigam, A., Oswalia, J., Kumar, S., & Gupta, R. (2024). Antimicrobial peptide-based strategies to overcome antimicrobial resistance. Archives of Microbiology, 206(10), 411.
9. Goles, M., Daza, A., Cabas-Mora, G., Sarmiento-Varón, L., Sepúlveda-Yañez, J., Anvari-Kazemabad, H., ... & Medina-Ortiz, D. (2024). Peptide-based drug discovery through artificial intelligence: towards an autonomous design of therapeutic peptides. Briefings in bioinformatics, 25(4), bbae275.
10. He, S., Stone, T. A., & Deber, C. M. (2021). Uncoupling amphipathicity and hydrophobicity: Role of charge clustering in membrane interactions of cationic antimicrobial peptides. Biochemistry, 60(34), 2586-2592.
11. Huynh, L., Velásquez, J., Rabara, R., Basu, S., Nguyen, H. B., & Gupta, G. (2021). Rational design of antimicrobial peptides targeting Gram-negative bacteria. Computational Biology and Chemistry, 92, 107475.
12. Islam, M. M., Asif, F., Zaman, S. U., Arnab, M. K. H., Rahman, M. M., & Hasan, M. (2023). Effect of charge on the antimicrobial activity of alpha-helical amphibian antimicrobial peptide. Current Research in Microbial Sciences, 4, 100182.
13. Li, C., Sutherland, D., Richter, A., Coombe, L., Yanai, A., Warren, R. L., ... & Birol, I. (2024). De novo synthetic antimicrobial peptide design with a recurrent neural network. Protein Science, 33(8), e5088.
14. Lin, C., Xiong, S., Cui, F., Zhang, Z., Shi, H., & Wei, L. (2025). Deep learning in antimicrobial peptide prediction. Journal of Chemical Information and Modeling, 65(14), 7373-7392.
15. Lin, T. T., Yang, L. Y., Lu, I. H., Cheng, W. C., Hsu, Z. R., Chen, S. H., & Lin, C. Y. (2021).

AI4AMP: an antimicrobial peptide predictor using physicochemical property-based encoding method and deep learning. Msystems, 6(6), e00299-21.

16. Malshikare, H., Priyakumar, U. D., Chatterjee, P., & Sengupta, D. (2026). Mechanistic principles of antimicrobial peptides uncovered by charge density-based machine learning. Chemical Communications, 62(13), 4067-4070.
17. Marczak, B., Bocian, A., & Łyskowski, A. (2026). MarLys AMP database - MLAMP_db (Version 3) [Data set]. Mendeley Data. https://doi.org/10.17632/w4hb5grjwb.3
18. Mondal, R. K., Sen, D., Arya, A., & Samanta, S. K. (2023). Developing anti-microbial peptide database version 1 to provide comprehensive and exhaustive resource of manually curated AMPs. Scientific Reports, 13(1), 17843.
19. Pandit, G., Chowdhury, N., Abdul Mohid, S., Bidkar, A. P., Bhunia, A., & Chatterjee, S. (2021). Effect of secondary structure and side chain length of hydrophobic amino acid residues on the antimicrobial activity and toxicity of 14-residue-long de novo AMPs. ChemMedChem, 16(2), 355-367.
20. Pirtskhalava, M., Amstrong, A. A., Grigolava, M., Chubinidze, M., Alimbarashvili, E., Vishnepolsky, B., ... & Tartakovsky, M. (2021). DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. Nucleic acids research, 49(D1), D288-D297.
21. Roy, K. (2020). Antimicrobial peptides (Version 1) [Data set]. Mendeley Data. https://doi.org/10.17632/37gbg3p3nr.1
22. Schifano, N. P., & Caputo, G. A. (2022). Investigation of the role of hydrophobic amino acids on the structure–activity relationship in the antimicrobial venom peptide ponericin L1. The Journal of membrane biology, 255(4), 537-551.
23. Sun, Z., Wang, X., Kuang, S., Song, J., & de la Fuente-Nunez, C. (2026). From generation to validation: Deep generative models for antimicrobial peptide discovery. Current Opinion in Chemical Biology, 93, 102685.
24. Szymczak, P., & Szczurek, E. (2023). Artificial intelligence-driven antimicrobial peptide discovery. Current Opinion in Structural Biology, 83, 102733.
25. Teimouri, H., Medvedeva, A., & Kolomeisky, A. B. (2023). Bacteria-specific feature selection for enhanced antimicrobial peptide activity predictions using machine-learning methods. Journal of Chemical Information and Modeling, 63(6), 1723-1733.
26. Van der Walt, M., Moller, D. S., Van Wyk, R. J., Ferguson, P. M., Hind, C. K., Clifford, M., ... & Gaspar, A. R. (2024). QSAR reveals decreased lipophilicity of polar residues determines the selectivity of antimicrobial peptide activity. ACS omega, 9(24), 26030-26049.
27. Wan, F., Wong, F., Collins, J. J., & de la Fuente-Nunez, C. (2024). Machine learning for antimicrobial peptide identification and design. Nature Reviews Bioengineering, 2(5), 392-407.
28. Wang, G. (2022). Unifying the classification of antimicrobial peptides in the antimicrobial peptide database. In Methods in enzymology (Vol. 663, pp. 1-18). Academic Press.
29. Wang, G. (2023). The antimicrobial peptide database is 20 years old: Recent developments and future directions. Protein Science, 32(10), e4778.
30. Wang, G., Vaisman, I. I., & Van Hoek, M. L. (2022). Machine learning prediction of antimicrobial peptides. In Computational peptide science: Methods and protocols (pp. 1-37). New York, NY: Springer US.
31. Wang, X. F., Tang, J. Y., Sun, J., Dorje, S., Sun, T. Q., Peng, B., ... & Wang, D. B. (2024). ProT-Diff: a modularized and efficient strategy for de novo generation of antimicrobial peptide sequences by integrating protein language and diffusion models. Advanced Science, 11(43), 2406305.