IJBB
International Journal of Biotechnology and Biomedicine
Vol. 3 No2; July 2026 https://doi.org/10.31674/ijbb.2025.v03i02.001

Integration of Protein Language Model Embeddings and Quantitative Structure-Activity Relationship Descriptors in a Gradient Boosting Framework for Antimicrobial Peptide Classification

*** Jianwei GONG[1]**

[1]Department of Biomedical Engineering, City University of Hong Kong | Department of Physics, City University of Hong Kong

****Corresponding Author's E-mail***: jianwgong2-c@my.cityu.edu.hk
https://orcid.org/0009-0002-2506-8952

Article received on 14th May 2026.
Revision received on 19th June 2026.
Accepted on 5th July 2026.

Abstract
In this work, we developed a hybrid eXtreme Gradient Boosting (XGBoost) framework for antimicrobial peptide (AMP) classification using a dataset of 20,968 sequences (2,001 AMPs). The framework combines embeddings from the ESM-2 protein language model with five quantitative structure-activity relationship (QSAR) features (net charge, hydrophobicity, length, positive residue count and hydrophobic residue ratio). Using five-fold nested cross-validation to avoid data leakage and scale_pos_weight to handle class imbalance, we conducted a systematic comparison of our hybrid model against three baselines: amino acid composition (AAC)-only, QSAR-only, and ESM-only. The hybrid model achieved an average AUC of 0.9137 and MCC of 0.6446, outperforming all three baselines. SHapley Additive exPlanations (SHAP) analysis identified sequence length as the most important feature, followed by several ESM-2 dimensions, with QSAR features playing a supporting role. This lightweight, interpretable framework offers a practical tool for computational AMP screening.

Keywords: Antimicrobial Peptides; ESM-2; QSAR Descriptors; XGBoost; Nested Cross-validation; SHAP Analysis.

1.0 Introduction
Antibiotic resistance has become a major global public health issue. The effectiveness of traditional antimicrobial drugs continues to decline, which makes clinical treatment of infections increasingly difficult. Antimicrobial peptides (AMPs) offer a promising alternative due to their broad-spectrum activity and low tendency to induce resistance, as seen with clinically used agents such as polymyxin B and colistin for treating multidrug-resistant Gram-negative infections (Bae *et al*., 2025; Yu *et al*., 2026). Developing efficient and accurate computational methods to screen for AMPs is therefore an important step toward accelerating their discovery and application.

Traditional experimental screening of AMPs is costly, low-throughput, and time-consuming, making it impractical for large-scale sequence screening. Machine learning-based prediction methods have helped address this problem, but current AMP classification models still have several limitations. Many traditional models rely on handcrafted features such as amino acid composition or simple QSAR descriptors, which have limited representational power (Bale *et al*., 2023; Bhatnagar *et al*., 2024; Bournez *et al*., 2023; Lertampaiporn *et al*., 2022; Mera-Banguero *et al*., 2024; Singh, 2022). Methods based solely on protein language models do not combine deep embedding features with classic physicochemical properties, thus missing the potential benefits of feature complementarity (Bin *et al*., 2025; Chen *et al*., 2026; Cordoves-Delgado and García-Jacas, 2024; Georgoulis *et al*., 2025; Lin *et al*., 2026). Moreover, many studies do not adopt a rigorous nested cross-validation strategy, which can lead to data leakage during hyperparameter tuning and result in biased performance estimates (Ali *et al*., 2026; Brizuela *et al*., 2025; García-González *et al*., 2025). In addition, systematic comparisons with multiple baseline models are often lacking, and few studies use interpretability methods to analyze feature contributions, making it unclear which factors actually drive model performance (Abbas *et al*., 2025; Jullapech, 2025).
On the data side, we collected natural AMPs from the DRAMP and APD databases as positive samples, and used sequences from the UniProt database as real negative samples. A robust FASTA parsing method was used to handle encoding errors and illegal characters. We then applied VSEARCH to remove redundant sequences, using 98% identity for positive samples and 90% for negative samples, to ensure a clean and reliable dataset. For feature extraction, we used the lightweight ESM-2_t6_8M_UR50D model to obtain deep sequence embeddings, and combined them with five core QSAR descriptors to build hybrid features that capture both sequence semantics and physicochemical properties. For model building and validation, we used XGBoost as the base classifier, performed hyperparameter optimization with Optuna, and adopted a five-fold nested cross-validation strategy with an inner hold-out set to completely avoid data leakage (Malshikare *et al*., 2026; Shaon *et al*., 2024).

2.0 Related Work

Computational methods for AMP prediction have evolved over the years, moving from traditional handcrafted feature engineering toward deep learning and protein language models, often in hybrid modeling frameworks. Different technical approaches have their own strengths in terms of feature representation, model performance, and practical applicability. Insights from previous studies have provided valuable guidance for the method design in this work.

2.1 Traditional Computational Methods for AMP Prediction

Early AMP classification models mainly relied on handcrafted features and classical machine learning algorithms. Among these, amino acid composition (AAC) and quantitative structure-activity relationship (QSAR) descriptors were the two most widely used types of features. AAC captures the frequency of each of the 20 standard amino acids in a sequence. It is easy to compute and generally applicable, making it a long-standing baseline feature for AMP identification (Teimouri *et al*., 2023; Negovetić *et al*., 2024). The frequency pi p_i pi of the i i i-th amino acid is calculated as

$$p_i = \frac{n_i}{L}$$

where n_i is the count of amino acid i and L is the sequence length. QSAR descriptors, on the other hand, focus on physicochemical properties closely related to antimicrobial activity, such as net charge, hydrophobicity, sequence length, proportion of positively charged residues, and proportion of hydrophobic residues. The net charge is computed as

$$Net\ Charge = n_K + n_R - n_D - n_E$$

where n_K,n_R,n_D,n_E are the counts of lysine, arginine, aspartic acid, and glutamic acid, respectively.

These features directly reflect the connection between sequence structure and function and perform reasonably well on small datasets (Mohkam *et al*., 2024). Traditional methods typically use shallow models like support vector machines or random forests. They are lightweight and interpretable, but their feature design limits them to capturing only surface-level sequence information. As a result, they struggle to extract deeper semantic or functional patterns, especially on more complex datasets.

2.2 Recent Advances Based On Protein Language Models

With the maturation of Transformer-based unsupervised pretraining, protein language models have opened new directions for deep feature extraction in AMP prediction. The ESM series, in particular, has been widely adopted for biological sequence analysis due to its strong sequence representation capabilities. ESM-2, a lightweight pretrained model, captures amino acid context, structural preferences, and functional associations from large-scale sequence data through self-supervised learning. It produces high-dimensional sequence embeddings without requiring handcrafted features. In AMP prediction, using ESM-2 embeddings alone has already shown better performance than traditional AAC or QSAR features (Zahid *et al*., 2026). Recent studies have further explored Transformer-based and multi-source protein language model frameworks for AMP design and prediction (Pikalyova *et al*., 2025; Salimi and Lee, 2026; Shoombuatong *et al*., 2025). While researchers continue to explore the use of ESM models for peptide prediction, most studies still rely on single deep embeddings. How to effectively combine ESM-2's deep semantic features with classic physicochemical descriptors to fully exploit their complementary value remains an open area of investigation.

2.3 Limitations of Existing Hybrid Models and Ongoing Research Directions

Hybrid models that integrate deep features with traditional physicochemical properties have become a popular research direction in AMP prediction. Many attempts have been made in this area, but several aspects of the modeling pipeline still need further standardization and improvement. In terms of model validation, the proper use of nested cross-validation remains uncommon in many AMP prediction studies. Some earlier works performed hyperparameter tuning and final evaluation on the same data split, which can lead to overly optimistic performance estimates due to data leakage. By adopting a strict five-fold nested cross-validation strategy with a separate inner hold-out set for tuning, our approach ensures that performance metrics better reflect real-world generalization (Brizuela *et al*., 2025; García-González *et al*., 2025). Regarding model optimization, automated hyperparameter search is still not fully mature; most studies still rely on manual tuning. How to make the best use of efficient optimization algorithms to unlock

model potential is an ongoing effort. In addition, class imbalance in AMP datasets, the lack of standardized baseline comparisons, and the absence of interpretability analysis are also important issues that the community is working to address (Abbas *et al*., 2025; Jullapech, 2025). Research is gradually moving toward more rigorous and transparent modeling practices. However, several critical gaps remain insufficiently addressed: the lack of effective complementarity between deep embeddings and classic physicochemical descriptors, the absence of rigorous nested cross-validation in many studies, and the limited use of systematic baseline comparisons combined with interpretability analysis.

2.4 Gaps Addressed by This Work

To address the above gaps, we developed a hybrid XGBoost framework that integrates lightweight ESM-2 embeddings with five core QSAR descriptors. Our framework standardizes the modeling pipeline with stratified redundancy removal, five-fold nested cross-validation, automated hyperparameter tuning via Optuna, and SHAP-based interpretability analysis—features that are still relatively rare in existing hybrid AMP prediction studies. On feature fusion, a standardized approach that combines lightweight ESM-2 embeddings with core QSAR descriptors has yet to be fully explored and validated. On modeling pipelines, studies that integrate robust data preprocessing, stratified redundancy removal, nested cross-validation, automated hyperparameter tuning, and class imbalance handling into a single workflow are still relatively rare. In experimental analysis, systematic comparisons of AAC, QSAR, ESM-2, and hybrid features under identical settings, combined with interpretability analysis of feature contributions, could be further enriched. Unlike many previous hybrid models that often rely on single data splits or lack systematic comparisons, our framework combines stratified redundancy removal, five-fold nested cross-validation with inner hold-out tuning, automated hyperparameter optimization, and SHAP interpretability analysis within one consistent pipeline. This design aims to deliver more reliable and transparent results for AMP classification.

3.0 Construction of the Hybrid ESM-2 and QSAR-driven XGBoost Framework for AMP Classification

This section describes the complete methodological design of our AMP classification framework, covering dataset preprocessing, hybrid feature extraction, model building and optimization, validation strategy, interpretability analysis, and evaluation metrics. All steps follow standard bioinformatics practices and conventional machine learning protocols, with settings aligned to our actual experimental environment.

3.1 Dataset Construction and Preprocessing

We built our experimental dataset using three public biological databases to ensure authenticity and generalizability. Positive samples (AMPs) were taken from the general AMP collection in the DRAMP database and the natural AMP collection in the APD database. Negative samples (non-AMPs) were taken from real protein sequences in the UniProt database. This selection reflects real-world scenarios in AMP screening.

To handle common issues in biological sequence analysis such as encoding errors and illegal characters, we designed a robust FASTA parsing pipeline. We read sequence files using UTF-8 encoding with error replacement mode, kept only the 20 standard amino acids (ACDEFGHIKLMNPQRSTVWY), and automatically filtered out non-standard characters. We

IJBB
International Journal of Biotechnology and Biomedicine
Vol. 3 No2; July 2026 https://doi.org/10.31674/ijbb.2025.v03i02.001

also applied length filtering to remove sequences shorter than 10 amino acids or longer than 100 amino acids. During the subsequent VSEARCH redundancy removal, sequences shorter than 32 amino acids were automatically discarded. As a result, the shortest sequence retained in the final dataset is 32 amino acids.

Sequence redundancy is a known cause of overestimated model performance. To minimize the risk of overestimated performance due to highly similar sequences appearing in both training and test sets, we performed redundancy removal separately for positive and negative samples using VSEARCH. For the positive AMP sequences collected from DRAMP and APD, we applied a high identity threshold of 98%. This relatively lenient threshold removes exact duplicates while preserving biologically similar variants, thereby retaining natural sequence diversity within the limited pool of known antimicrobial peptides. In contrast, the much larger negative sequence set from UniProt was filtered at a stricter 90% identity threshold to more aggressively reduce redundancy. This differential strategy helps prevent the model from exploiting trivial sequence similarities and leads to more realistic performance estimates. Finally, we limited the number of negative samples to under 25,000 to keep the overall dataset size practical for efficient model training. After full preprocessing, the dataset was stored in a standardized CSV format containing three core fields: sequence ID, amino acid sequence, and class label.

Figure 1 shows the overall dataset preprocessing and redundancy removal pipeline. The final dataset statistics are summarized in Table 1, and the distributions of QSAR features are presented in Table 2.

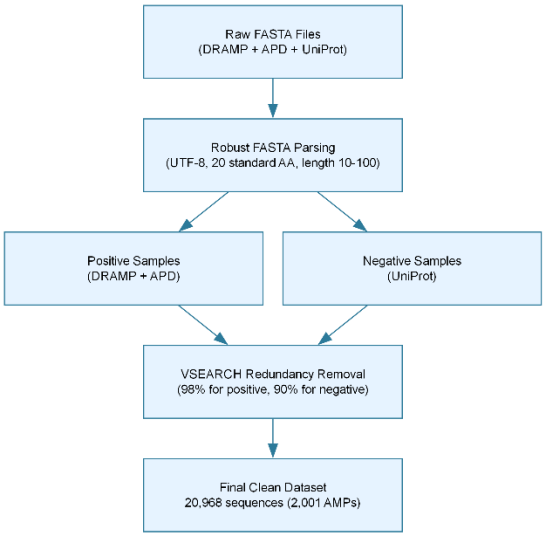

Figure 1. Schematic diagram of the dataset preprocessing and redundancy removal pipeline.

Table 1
Summary statistics of the dataset

Dataset	Number of Samples	Positive Sample Proportion
Training Set	16,774	0.0954
Test Set	4,194	0.0954
Total Dataset	20,968	0.0954

IJBB
International Journal of Biotechnology and Biomedicine
Vol. 3 No2; July 2026 https://doi.org/10.31674/ijbb.2025.v03i02.001

Table 2
Statistics of QSAR features

Statistic	Net Charge	Hydrophobicity	Length	Positive Residue Count	Hydrophobic Residue Ratio
Count	20,968	20,968	20,968	20,968	20,968
Mean	1.3500	-0.1320	53.0710	1.8800	0.0555
Std	2.3190	0.2680	10.6850	2.7670	0.0770
Min	-6.0000	-1.9860	32.0000	0.0000	0.0000
25%	0.0000	-0.1810	45.0000	0.0000	0.0000
50%	0.0000	-0.0140	53.0000	1.0000	0.0220
75%	2.0000	0.0000	62.0000	3.0000	0.0790
Max	17.0000	1.0410	93.0000	18.0000	0.5680

3.2 Hybrid Feature Extraction

We adopted a strategy that combines deep semantic features from ESM-2 with traditional physicochemical properties from QSAR descriptors. This integration is biologically meaningful because AMP activity is governed by both sequence context (e.g., evolutionary patterns and residue interactions captured by ESM-2 embeddings) and key physicochemical properties such as net charge and hydrophobicity, which directly influence membrane interaction and antimicrobial mechanisms. For example, while ESM-2 may encode contextual information about residue positions that correlate with function, the explicit net charge and hydrophobic residue ratio features provide direct, interpretable signals of electrostatic attraction and membrane insertion—properties known to be critical for AMP activity but not always explicitly represented in learned embeddings alone. By concatenating the 320-dimensional ESM-2 embeddings with the five QSAR features, the hybrid representation captures both implicit semantic patterns and explicit biological attributes. For deep feature extraction, we used the lightweight protein language model ESM-2_t6_8M_UR50D. This model has a small number of parameters and fast inference speed, making it suitable for conventional GPU environments. We set the maximum sequence length to 60 because most antimicrobial peptides are relatively short, with a mean length of 53 in our dataset. This length covers the majority of relevant sequence information while keeping computational cost manageable on standard GPU hardware. Sequences longer than 60 amino acids were truncated to the first 60 residues, while shorter sequences were padded with zeros to a uniform length of 60. Although truncation may cause some loss of information in the C-terminal region of longer sequences, the impact is limited in our dataset because the average sequence length is 53 residues. This design choice prioritizes computational efficiency and model compatibility while maintaining good coverage of typical AMP sequences. However, it may slightly reduce generalizability when applied to longer peptides outside the current length distribution.

For traditional physicochemical features, we selected five core QSAR descriptors that are highly relevant to antimicrobial activity. These features are derived from known amino acid properties, have clear biological meanings, and are computationally efficient. They help complement the biological interpretability of the deep embeddings. The final hybrid feature vector was obtained by concatenating the 320 ESM-2 features and the 5 QSAR features, resulting in a 325-dimensional

Vol. 3 No2; July 2026 https://doi.org/10.31674/ijbb.2025.v03i02.001

vector for classification. Figure 2 illustrates the architecture of the hybrid feature extraction pipeline.

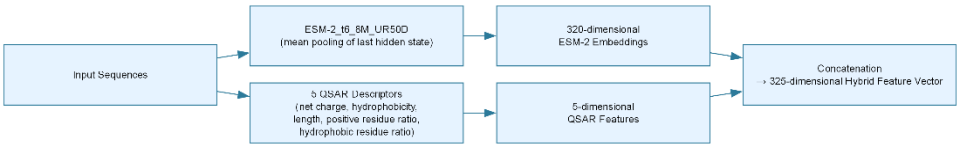

Figure 2. Architecture of the hybrid feature extraction pipeline.

3.3 Model Architecture and Hyperparameter Optimization

We selected XGBoost as the base classifier for the hybrid model for several reasons. XGBoost efficiently handles high-dimensional hybrid features (325 dimensions in this study), offers strong resistance to overfitting through regularization, and supports GPU acceleration, which is advantageous for moderately large datasets. Additionally, its built-in support for class imbalance via the scale_pos_weight parameter makes it particularly suitable for AMP classification tasks where positive samples are scarce. These characteristics collectively provide a good balance of predictive performance, computational efficiency, and interpretability.

To fully explore model performance, we used Optuna for automated hyperparameter optimization, with the validation AUC as the optimization objective. Key hyperparameters searched included learning rate, subsample ratio, column sampling ratio, tree depth, number of boosting rounds, L1 regularization coefficient, and L2 regularization coefficient. We performed 20 optimization trials because additional trials showed diminishing returns in performance improvement while incurring higher computational cost on the available GPU resources. During training, we enabled the histogram-based algorithm and GPU acceleration on an RTX 3060 GPU to speed up training and hyperparameter search. After optimization, the best hyperparameter combination was retained for training the final classification model. Figure 3 presents the workflow of XGBoost model training and Optuna hyperparameter optimization. The optimal hyperparameters are summarized in Table 3.

IJBB
International Journal of Biotechnology and Biomedicine
Vol. 3 No2; July 2026 https://doi.org/10.31674/ijbb.2025.v03i02.001

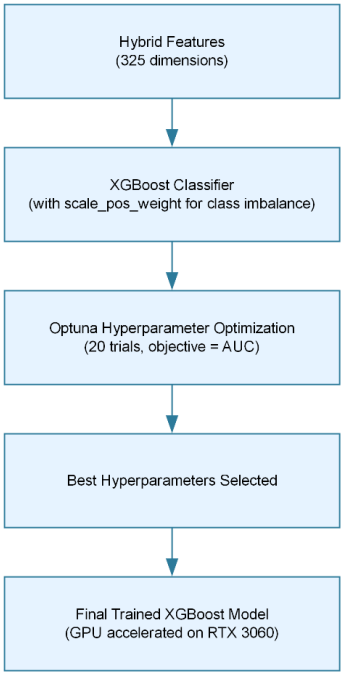

Figure 3. Workflow of XGBoost model training and Optuna hyperparameter optimization.

Table 3
Optimal hyperparameters of the model

Parameter	Value
n_estimators	384
max_depth	7.0000
learning_rate	0.0786
subsample	0.9055
colsample_bytree	0.8576

3.4 Nested Cross-Validation Strategy

To avoid data leakage during hyperparameter optimization and to ensure realistic and generalizable performance evaluation, we adopted a five-fold nested cross-validation strategy that strictly separates the data used for hyperparameter tuning from that used for model evaluation. The strategy consists of an outer loop and an inner loop. In the outer loop, the entire dataset is split into five folds while preserving class distribution. In each iteration, one fold is held out as the independent test set, and the remaining four folds serve as the outer training set. This gives five rounds of independent model evaluation. In the inner loop, the outer training set is further split at a 3:1 ratio, where 25% of the data is set aside as a hold-out validation set specifically for evaluating candidate hyperparameters during optimization. Once hyperparameter optimization is complete, the final model is trained on the full outer training set. This design completely prevents data leakage from the validation process and yields performance estimates that better reflect real-world applications. Figure 4 shows the schematic diagram of the five-fold nested cross-validation strategy.

IJBB
International Journal of Biotechnology and Biomedicine
Vol. 3 No2; July 2026 https://doi.org/10.31674/ijbb.2025.v03i02.001

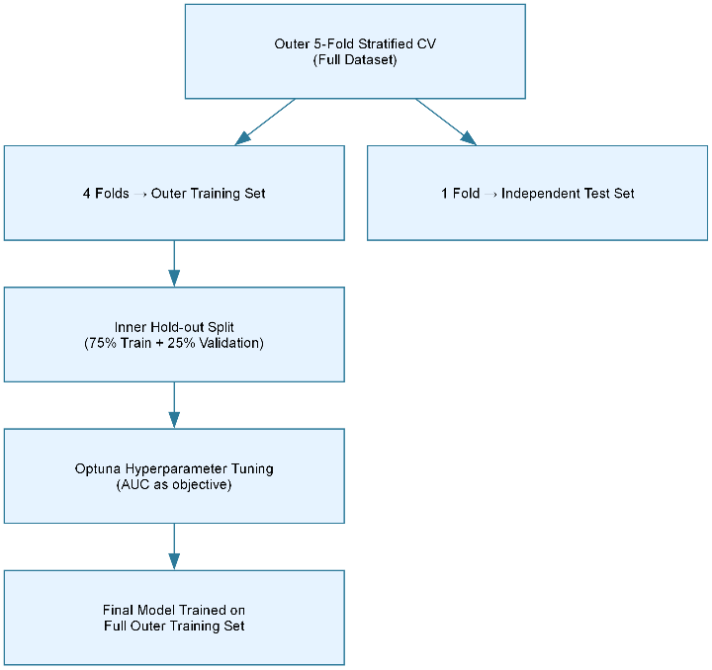

Figure 4. Schematic diagram of the five-fold nested cross-validation strategy.

3.5 Interpretability Analysis

To open the black box of the machine learning model, we used the SHAP (SHapley Additive exPlanations) algorithm for interpretability analysis. SHAP is grounded in game theory and can quantify each feature's contribution to the classification outcome. It supports both global feature importance analysis and local explanation of individual predictions. In our experiments, we selected representative samples from the test set, computed feature contributions using a SHAP explainer, and generated a global feature importance summary. We also reported the top 10 features by mean absolute SHAP value, providing insight into the key factors driving AMP classification.

3.6 Evaluation Metrics

Given the imbalanced nature of the AMP classification task, we used multiple evaluation metrics to obtain a comprehensive view of model performance and to avoid relying on any single measure. The core metrics included accuracy, area under the ROC curve (AUC), Matthews correlation coefficient (MCC), precision, recall, and F1 score. Among these, AUC and MCC are more sensitive to imbalanced data and better reflect the model's overall classification ability; they served as the primary metrics for optimization and comparison. The other metrics were used as auxiliary measures to evaluate different aspects of model performance, forming a complete and objective evaluation system.

4.0 Results and Discussion

This section presents the full experimental results of our AMP classification framework based on the methods described in Section 3. Using a combination of quantitative metrics and visualizations, we demonstrate the performance advantages of the hybrid model, interpret the contribution of

different features, and discuss the strengths and limitations of the approach. All results are derived from actual experimental runs and reflect the real performance of the model.

4.1 Dataset Summary

For model training and evaluation, we first split the entire dataset into training and test sets using stratified sampling with an 8:2 ratio, preserving the class distribution in both subsets. All subsequent hyperparameter optimization and model selection were performed exclusively on the training set using five-fold nested cross-validation. The independent test set was only used for final performance evaluation of the optimal model. This avoids bias in performance estimation due to uneven data splitting.

4.2 Comparative Results from Five-Fold Cross-Validation

To evaluate the performance advantage of combining ESM 2 with QSAR features, we set up four baseline models for a fair comparison: an AAC-only model using only amino acid composition features, a QSAR-only model using only the five QSAR descriptors, an ESM-only model using only ESM 2 deep embeddings, and our proposed hybrid model. All models used the same XGBoost architecture, the same Optuna hyperparameter optimization procedure, and the same five-fold nested cross-validation strategy. This minimizes differences in experimental settings and ensures objective comparison.

The average results from five-fold cross-validation show that the hybrid model outperforms all three baselines across all core evaluation metrics, achieving the best overall performance. Among the baseline models, the ESM-only model performed significantly better than the traditional feature-based models, thanks to the deep sequence representation power of the protein language model. The AAC-only model came next, showing that basic sequence composition already provides useful information for classification. The QSAR-only model, relying only on five handcrafted physicochemical features, showed relatively limited performance. This is likely because these traditional descriptors, while biologically meaningful, have restricted representational capacity and cannot fully capture the intricate sequence patterns and contextual relationships present in antimicrobial peptides. In contrast, the ESM-2 embeddings provide richer, learned representations that better handle the complexity of the task. These results clearly demonstrate that ESM 2 deep embeddings and QSAR physicochemical features have strong complementary effects. Their combination effectively enhances feature representation and improves both classification accuracy and generalization. The average performance of the four models across five-fold cross-validation is summarized in Table 4. Figure 5 presents the AUC comparison across the four models. Figure 6 shows the MCC comparison. The per-fold AUC values across the five folds are illustrated in Figure 7. A comprehensive multi-metric performance comparison is provided in Figure 8.

To visualize the performance differences, we generated several plots from the cross-validation results. The AUC comparison bar chart clearly shows that the hybrid model has significantly better discriminative ability than the baselines. The MCC comparison bar chart indicates that the hybrid model also performs better on imbalanced data. The per-fold AUC line plot shows very little fluctuation across the five folds for all four models, confirming that the nested cross-validation strategy effectively reduces the risk of data leakage and ensures stable results. The multi-metric comparison bar chart further demonstrates the overall advantage of the hybrid model across accuracy, precision, recall, and other metrics. It should be noted that while the hybrid model

IJBB
International Journal of Biotechnology and Biomedicine
Vol. 3 No2; July 2026 https://doi.org/10.31674/ijbb.2025.v03i02.001

outperformed the ESM-only baseline, the improvement in AUC and MCC was relatively modest. This suggests that ESM-2 embeddings already capture a substantial amount of relevant information for AMP classification. The addition of the five QSAR features provides complementary physicochemical signals but contributes only incremental gains. Possible reasons include the limited number of QSAR descriptors used and potential redundancy between some ESM-2 dimensions and handcrafted features. Future work could explore more advanced feature fusion strategies, such as attention-based integration or learned feature weighting, and incorporate a broader set of physicochemical or structural descriptors to further enhance the hybrid model's performance.

Table 4
Average performance comparison of the four models over five-fold cross-validation

Model	Accuracy	AUC	MCC	Precision	Recall	F1
AAC-only	0.9220	0.8898	0.5915	0.5846	0.6902	0.6304
QSAR-only	0.8733	0.8406	0.4499	0.4028	0.6582	0.4989
ESM-only	0.9343	0.8943	0.6250	0.6515	0.6716	0.6611
Hybrid	0.9363	0.9137	0.6446	0.6760	0.6817	0.6784

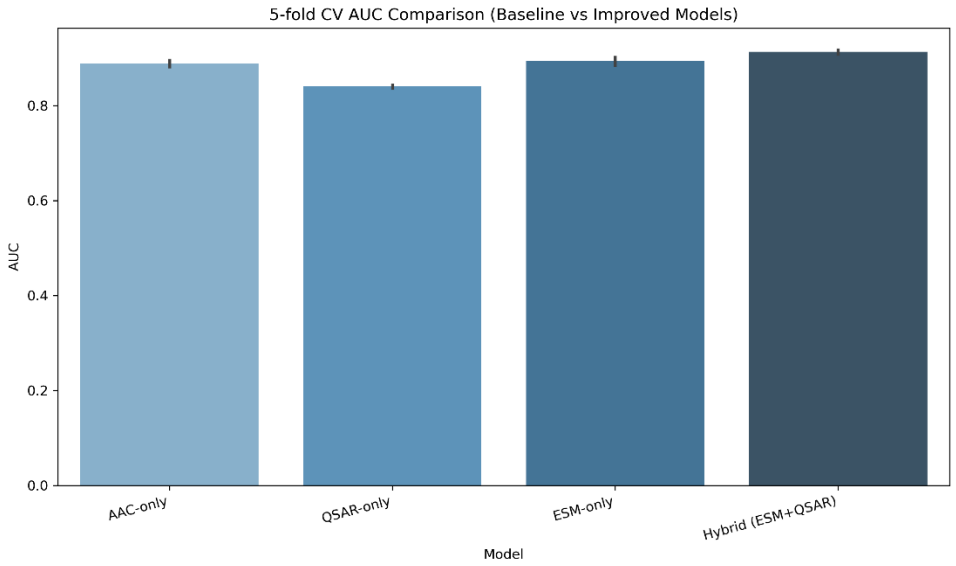

Figure 5. Bar chart comparing AUC values across the four models in five-fold cross-validation

IJBB
International Journal of Biotechnology and Biomedicine
Vol. 3 No2; July 2026 https://doi.org/10.31674/ijbb.2025.v03i02.001

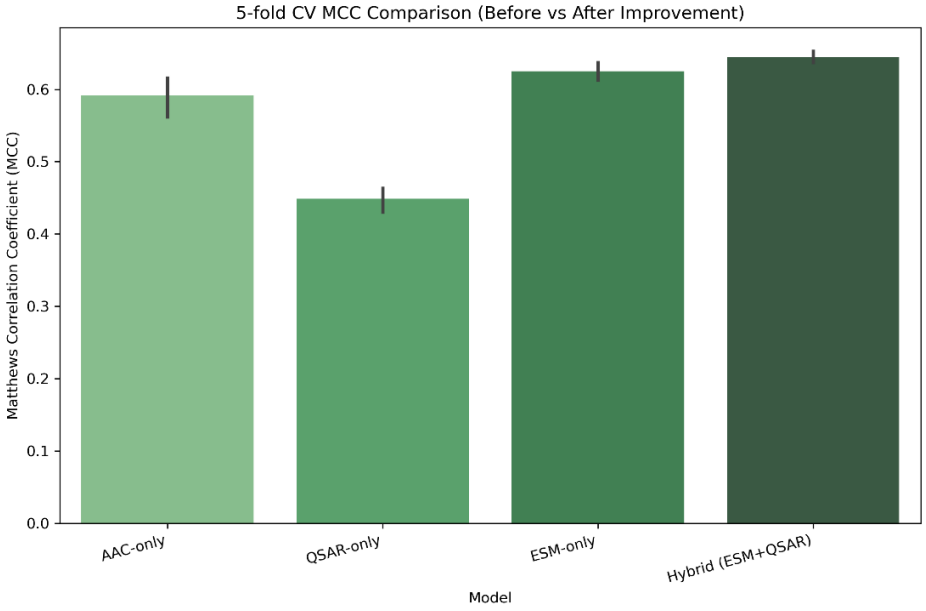

Figure 6. Bar chart comparing MCC values across the four models

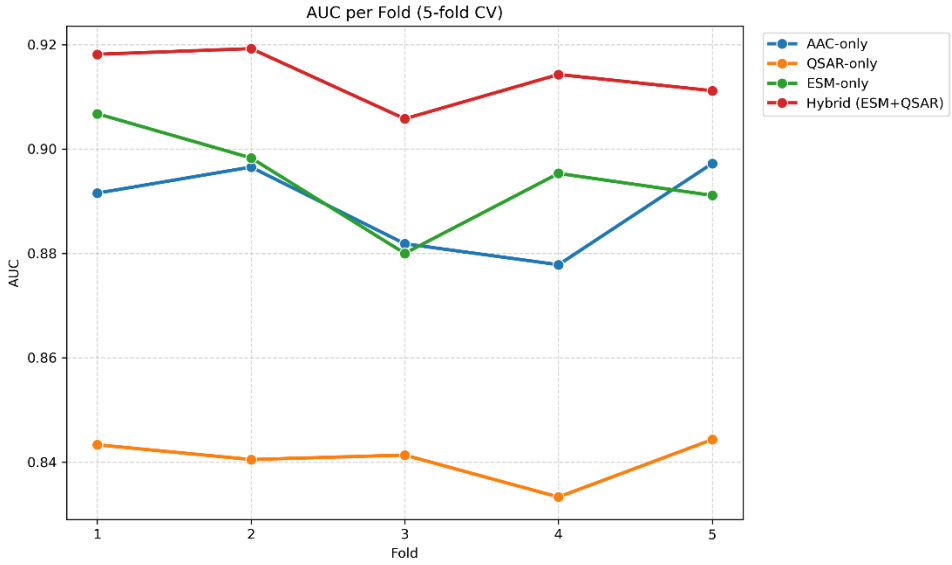

Figure 7. Line plot showing AUC per fold for each model

IJBB
International Journal of Biotechnology and Biomedicine
Vol. 3 No2; July 2026 https://doi.org/10.31674/ijbb.2025.v03i02.001

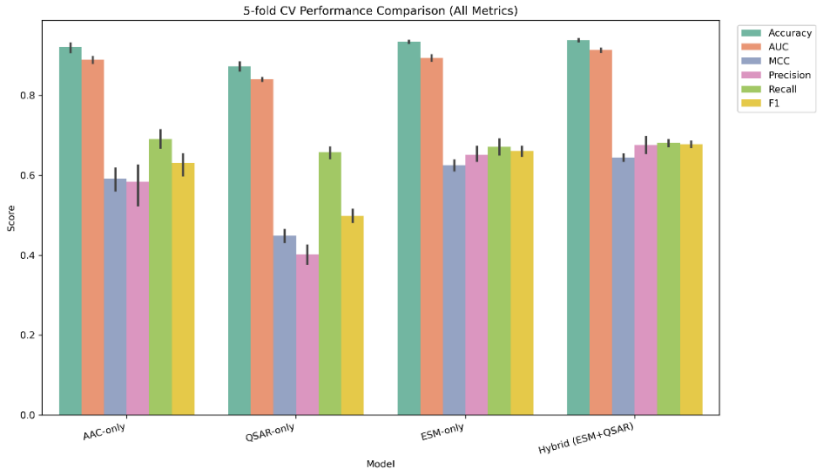

Figure 8. Multi-metric performance comparison across models

4.3 Detailed Performance of the Optimal Hybrid Model

Using the best hyperparameter combination selected from the five-fold cross-validation, we trained the final hybrid model and evaluated it on the independent test set. The model performed robustly on the imbalanced AMP classification task. The ROC curve of the final model is shown in Figure 9. The precision-recall curve is presented in Figure 10. The confusion matrix on the test set is displayed in Figure 11. The distribution of predicted probabilities is shown in Figure 12. Detailed performance metrics are summarized in Table 5, the confusion matrix counts are provided in Table 6, and the predicted probability statistics are listed in Table 7. We observed a modest drop in MCC from the cross-validation average (0.6446) to the independent test set (0.6221). Such variation is expected when moving from averaged cross-validation results to a single held-out test set and can arise from random partitioning effects or minor differences in data distribution. The relatively small difference supports the overall stability and robustness of the model. To fully examine its classification performance, we carried out visualizations from three perspectives: ROC curve, precision recall curve, confusion matrix, and prediction probability distribution.

The ROC curve shows that the model's curve deviates substantially from the diagonal line of random guessing, indicating an excellent ability to distinguish between positive and negative samples. The precision-recall curve maintains a high precision level even at high recall, which is particularly useful for AMP screening where identifying true positive candidates is a priority. The confusion matrix provides a detailed view of classification results: the model achieves high accuracy for negative samples, while the use of the scale_pos_weight parameter effectively improves recall for positive samples, mitigating the bias caused by class imbalance. The prediction probability distribution on the test set shows a clear separation between the predicted probabilities for positive and negative samples, with no large ambiguous region. This suggests that the model has learned a stable and effective decision boundary, and its predictions are reliable.

IJBB
International Journal of Biotechnology and Biomedicine
Vol. 3 No2; July 2026 https://doi.org/10.31674/ijbb.2025.v03i02.001

Table 5
Core performance metrics of the optimal hybrid model on the test set

Metric	Value
Accuracy	0.9423
AUC	0.9139
MCC	0.6221

Table 6
Confusion matrix counts on the test set

	Predicted Negative	Predicted Positive
True Negative	3,739	55
True Positive	187	213

Table 7
Statistics of predicted probabilities

Statistic	Probability
Count	4,194
Mean	0.0840
Std	0.2143
Min	0.0001
25%	0.0026
50%	0.0097
75%	0.0371
Max	0.9966

IJBB
International Journal of Biotechnology and Biomedicine
Vol. 3 No2; July 2026 https://doi.org/10.31674/ijbb.2025.v03i02.001

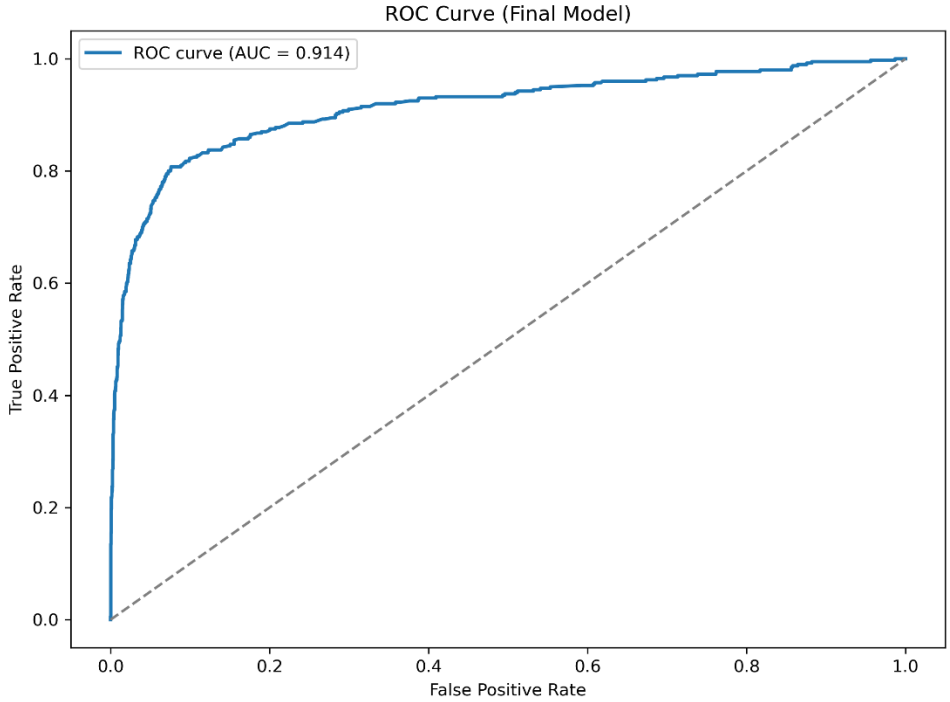

Figure 9. ROC curve of the optimal model

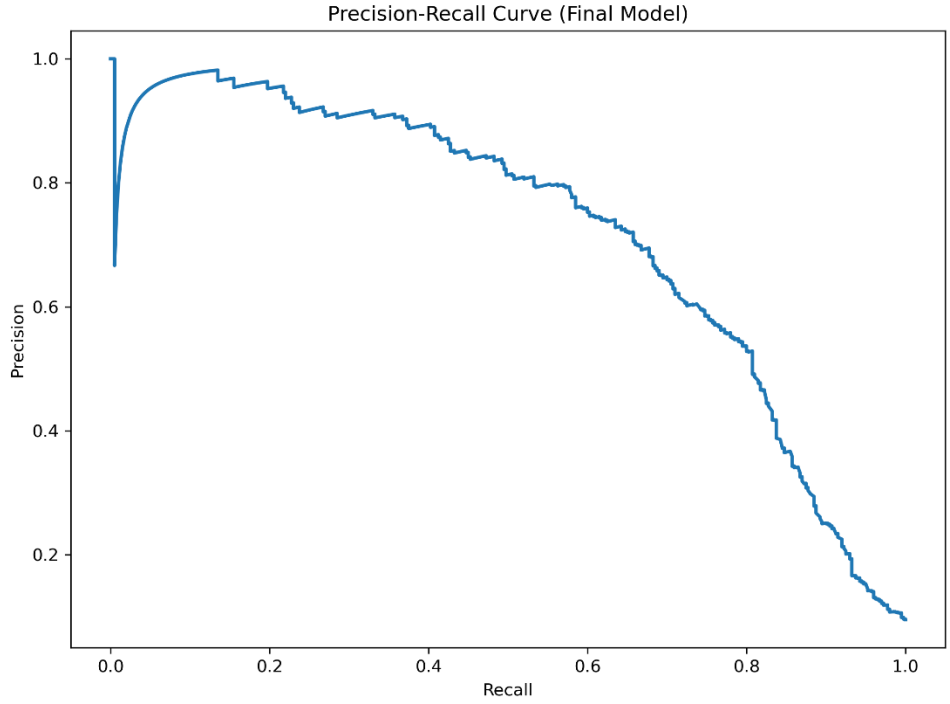

Figure 10. Precision-recall curve of the optimal model

IJBB
International Journal of Biotechnology and Biomedicine
Vol. 3 No2; July 2026 https://doi.org/10.31674/ijbb.2025.v03i02.001

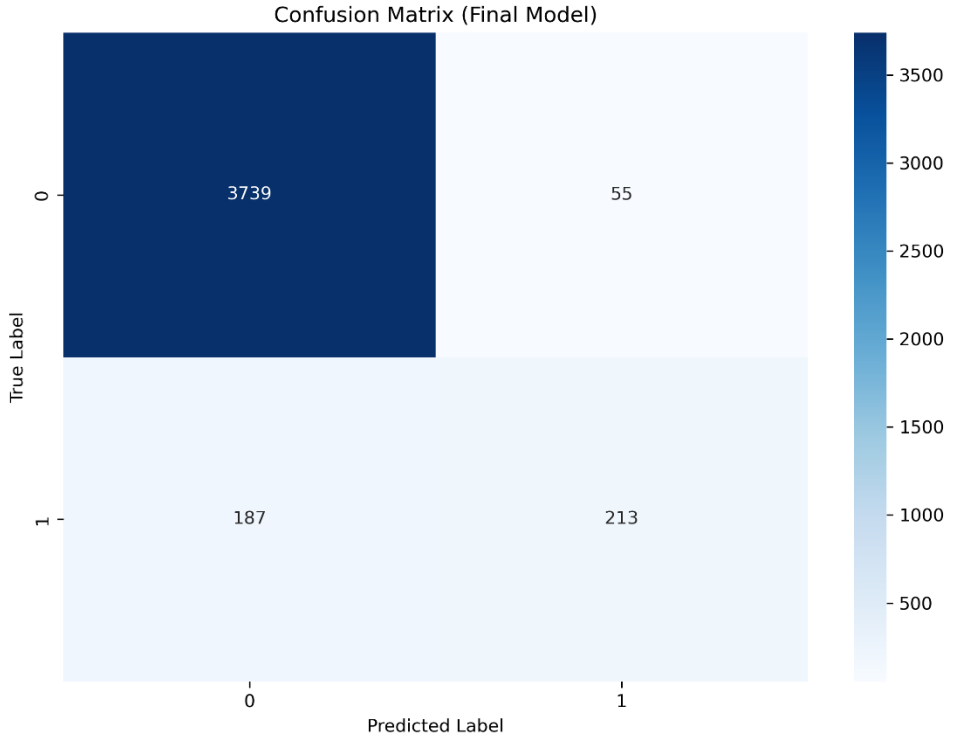

Figure 11. Confusion matrix heatmap of the optimal model

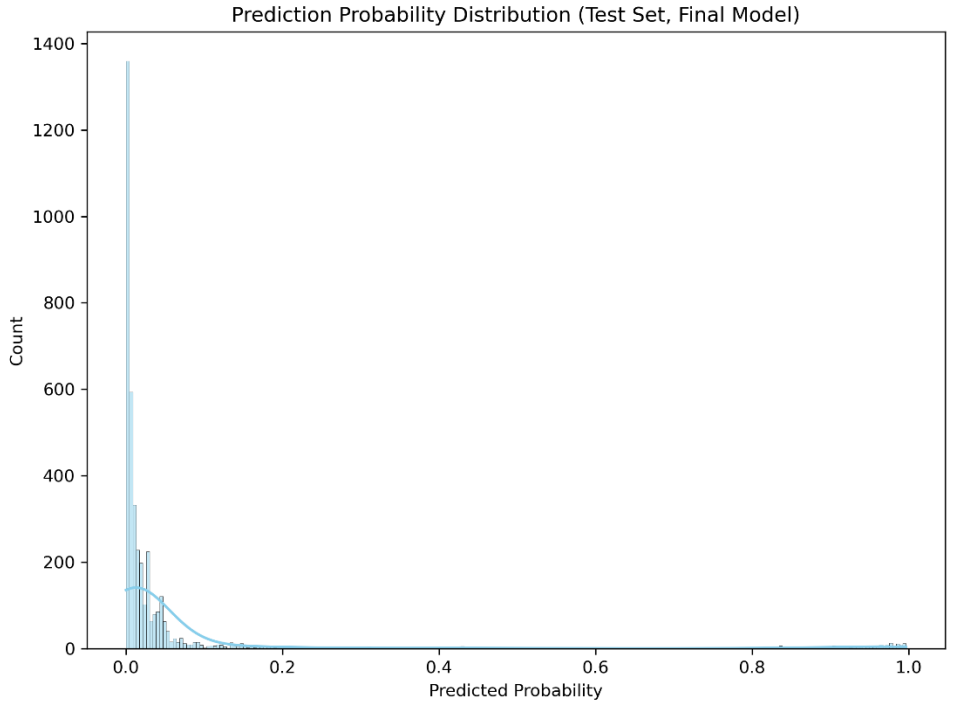

Figure 12. Histogram of prediction probability distribution on the test set

IJBB
International Journal of Biotechnology and Biomedicine
Vol. 3 No2; July 2026 https://doi.org/10.31674/ijbb.2025.v03i02.001

4.4 SHAP Interpretability Analysis

To uncover the decision mechanism of the hybrid model, we performed a global interpretability analysis using the SHAP algorithm. By quantifying feature contributions, we identified the key factors influencing classification results. The analysis shows that sequence length is the most important feature for AMP classification, with a contribution far greater than that of other features. Table 8 lists the top 10 features ranked by mean absolute SHAP value, providing a quantitative ranking of overall feature contributions. Figure 13 presents the global SHAP summary plot, which visualizes the direction and magnitude of each feature's impact on model predictions. Figure 14 shows the top 20 features by XGBoost built-in gain importance, which cross-validates the core feature conclusions from the SHAP analysis. This finding is consistent with the biological observation that natural AMPs are typically between 10 and 100 amino acids in length. In our final cleaned dataset after redundancy removal, the shortest sequence retained is 32 amino acids. Following sequence length, several ESM-2 embedding dimensions appear as important contributors. Although individual dimensions in ESM-2 embeddings do not correspond to simple, human-interpretable biological properties, their collective importance indicates that the model relies on rich, high-dimensional representations learned from large-scale evolutionary data. These representations are known to encode complex contextual and structural patterns that are relevant to protein function, including antimicrobial activity. Although QSAR features such as net charge and hydrophobicity did not enter the top-10 list, they still play a supporting role in model decisions. These explicit physicochemical descriptors provide direct, biologically interpretable signals that are known to influence key mechanisms of antimicrobial activity, such as membrane interaction and disruption. In this way, they complement the high-dimensional, implicit patterns captured by the ESM-2 embeddings, which primarily encode contextual and evolutionary information. This combination helps improve both the predictive performance and the biological transparency of the hybrid model.

The SHAP summary plot clearly shows the positive and negative contributions of each feature to the classification outcome. The model's decision logic aligns well with the known structure-activity relationships of AMPs, rather than being a simple numerical fit. These interpretability results not only increase the trustworthiness of the model but also provide quantitative guidance for AMP sequence screening, structural optimization, and rational design, thereby enhancing the practical value of computational predictions for downstream experimental studies.

IJBB
International Journal of Biotechnology and Biomedicine
Vol. 3 No2; July 2026 https://doi.org/10.31674/ijbb.2025.v03i02.001

Table 8
Top 10 features by mean absolute SHAP value

Feature	Mean|SHAP|
Length	0.8397
ESM_242	0.1627
ESM_131	0.1568
ESM_217	0.1298
ESM_87	0.1185
ESM_199	0.1119
ESM_55	0.1119
ESM_305	0.0962
ESM_193	0.0949
ESM_308	0.0934

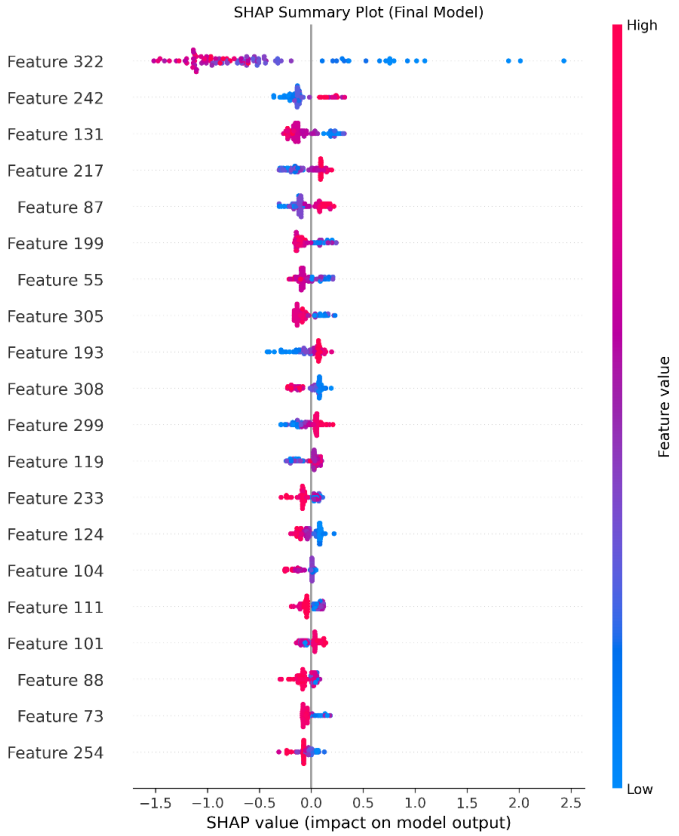

Figure 13. SHAP global feature importance summary plot

IJBB
International Journal of Biotechnology and Biomedicine
Vol. 3 No2; July 2026 https://doi.org/10.31674/ijbb.2025.v03i02.001

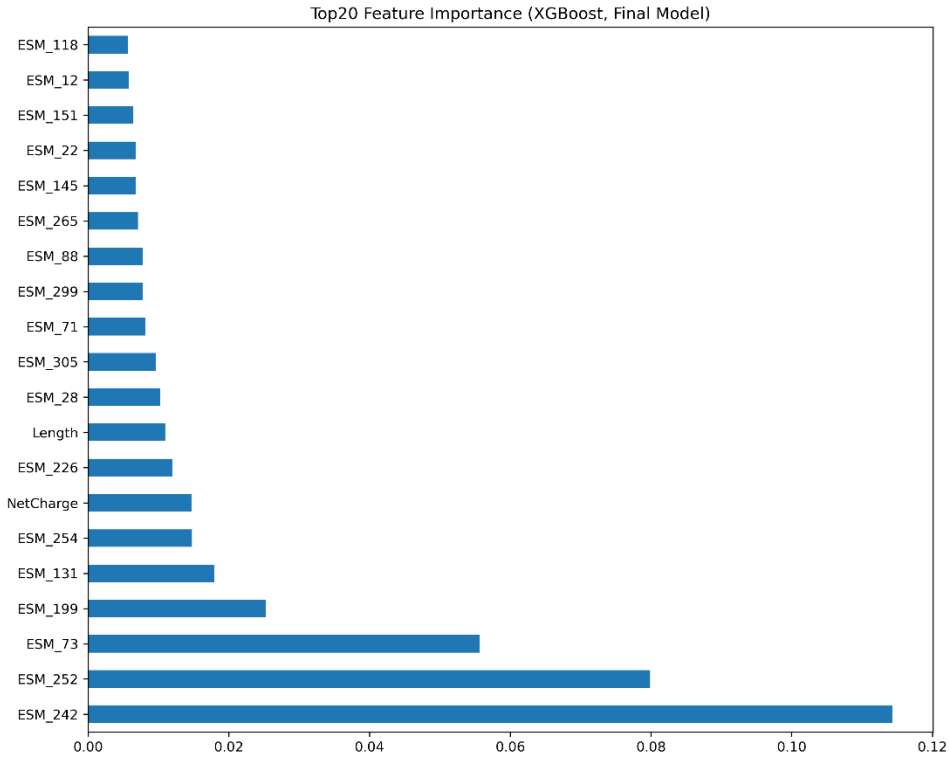

Figure 14. XGBoost feature importance bar chart (top-20)

4.5 Strengths, Practical Value, and Limitations

Compared with traditional AMP classification methods, our hybrid framework offers several practical advantages from an engineering perspective. First, it uses a lightweight ESM-2 model combined with GPU-accelerated XGBoost, resulting in relatively low computational cost that allows the pipeline to run efficiently on standard consumer-grade hardware. Second, the framework integrates data preprocessing, model optimization, and interpretability analysis into a standardized and reproducible workflow. These characteristics make the approach more accessible and practical for real-world AMP screening applications.

Nevertheless, this study has several limitations. First, although the dataset was constructed from three public databases with strict redundancy removal, it still lacks species-specific and synthetic AMP sequences. Further validation on more diverse independent datasets would be necessary to comprehensively evaluate the model's generalizability. Second, this work is purely computational; we have not performed wet-lab experiments to measure the activity of candidate AMPs predicted by the model. Experimental validation is needed to confirm the biological relevance of the predictions. Third, we only used five core QSAR features; incorporating additional physicochemical or structural descriptors could further improve performance. Fourth, we only used the lightweight ESM-2 model and did not explore larger protein language models. These limitations point to clear directions for future work, including expanding the dataset, incorporating experimental validation, enriching the feature set, and upgrading to larger pretrained models.

Vol. 3 No2; July 2026 https://doi.org/10.31674/ijbb.2025.v03i02.001

5.0 Conclusion and Future Work

In this study, we successfully constructed and validated a hybrid XGBoost framework for antimicrobial peptide classification by integrating ESM-2 protein language model embeddings with five QSAR physicochemical descriptors. Through standardized data preprocessing, rigorous nested cross-validation, systematic baseline comparisons, and SHAP-based interpretability analysis, our framework demonstrated clear advantages in classification accuracy, generalization, and interpretability, providing a practical and reusable computational solution for AMP screening. The experimental results show that deep sequence embeddings from ESM-2 and classic QSAR features have strong complementary effects. Their combination captures both the sequence semantics and biological properties of AMPs, leading to improved classification accuracy. The five-fold nested cross-validation strategy prevents data leakage during hyperparameter optimization, and together with automated hyperparameter search and class imbalance handling, ensures reliable performance estimates. SHAP analysis reveals that sequence length is the most important feature, followed by several ESM 2 embedding dimensions. While QSAR features such as net charge and hydrophobicity do not appear in the top 10, they still contribute as supporting factors, and the model's decisions align well with known biological patterns of AMPs. Overall, the framework is lightweight, efficient, and interpretable, offering a standardized and reusable computational solution for AMP screening.

Several directions can be pursued to further improve and extend this framework. First, the dataset can be expanded with more diverse, multi-species AMP sequences, and an independent external test set can be constructed to better assess generalizability. Second, computational predictions can be combined with wet lab experiments to validate the activity of candidate AMPs identified by the model, thereby increasing the practical impact of the approach. Third, the feature set can be enriched by including additional physicochemical and structural descriptors, and larger ESM pretrained models can be explored to unlock more powerful sequence representations. Fourth, the optimal model can be deployed as an online prediction tool with a lightweight web interface, making it easy for other researchers to use and promoting the wider adoption of computational methods in AMP development (Bae *et al*., 2025; Brizuela *et al*., 2025; Yu *et al*., 2026).

While this framework shows promising performance with good interpretability, some limitations should be noted. The adoption of a compact ESM-2 model and GPU-accelerated training results in relatively modest computational requirements, making the method accessible on standard hardware. However, the current model was trained solely on natural antimicrobial peptides. Its ability to generalize to synthetic or de novo designed peptides, which may have different sequence characteristics, has not been tested. Furthermore, although nested cross-validation and regularization techniques were used to reduce the risk of overfitting, this risk cannot be completely eliminated, particularly when the model encounters sequences that differ significantly from the training data in length or composition. Future work will aim to evaluate the framework on synthetic AMP datasets and explore ways to further improve its robustness.

Declarations

Since the author's native language is not English, the author used large language models such as DeepSeek and Grok to improve English expression.

Conflicts of Interest Declaration

All authors declare that they have no conflicts of interest.

References

Abbas, Munawar, *et al*. (2025) ABP-Xplorer: A Machine Learning Approach for Prediction of Antibacterial Peptides Targeting Mycobacterium abscessus-tRNA-Methyltransferase (TrmD). Journal of Chemical Information and Modeling, 65.11: 5456-5468.

Ali, Farman, *et al*. (2026) A generative explainable model for antimicrobial peptide prediction using bidirectional temporal convolutional neural network. Scientific Reports.

Bae, Daehun, *et al*. (2025) AI-guided discovery and optimization of antimicrobial peptides through species-aware language model. Briefings in Bioinformatics, 26.4: bbaf343.

Bale, Ashwin; Dutta, Arnab; Mitra, Debirupa. (2023) Combined charge and hydrophobicity-guided screening of antibacterial peptides: two-level approach to predict antibacterial activity and efficacy. Amino Acids, 55.7: 853-867.

Bhatnagar, Pranshul, *et al*. (2024) Predicting antibacterial activity, efficacy, and hemotoxicity of peptides using an explainable machine learning framework. Process Biochemistry, 145: 163-174.

Bin, Yannan, *et al*. (2025) Pepxml: ESM2-based extreme multilabel classification of pathogen-targeted antimicrobial peptides. Briefings in Bioinformatics, 26.5: bbaf548.

Bournez, Colin, *et al*. (2023) CalcAMP: A new machine learning model for the accurate prediction of antimicrobial activity of peptides. Antibiotics, 12.4: 725.

Brizuela, Carlos A., *et al*. (2025) AI methods for antimicrobial peptides: progress and challenges. Microbial Biotechnology, 18.1: e70072.

Chen, QingWei, *et al*. (2026) iAMP-SeE: an antimicrobial peptide recognition model based on ESM2 feature extraction and hybrid attention mechanisms. PeerJ, 14: e20978.

Cordoves-Delgado, Greneter; García-Jacas, César R. (2024) Predicting antimicrobial peptides using ESMFold-predicted structures and ESM-2-based amino acid features with graph deep learning. Journal of Chemical Information and Modeling, 64.10: 4310-4321.

García-González, Luis A., *et al*. (2025) Optimal Descriptor Subset Search via Chemical Information and Target Activity-Guided Algorithm for Antimicrobial Peptide Prediction. Journal of Chemical Information and Modeling, 65.13: 6621-6631.

Georgoulis, Elias; Zervou, Michaela Areti; Pantazis, Yannis. (2025) Transfer learning on protein language models improves antimicrobial peptide classification. Scientific Reports, 15.1: 37456.

Jullapech, Nawisa. (2025) Development and application of Machine Learning classification methods: optimal feature identification and prediction of antimicrobial peptides and other soft matter systems. 2025. PhD Thesis. University of Reading.

Lertampaiporn, Supatcha, *et al*. (2022) Ensemble-AHTPpred: a robust ensemble machine learning model integrated with a new composite feature for identifying antihypertensive peptides. Frontiers in Genetics, 2022, 13: 883766.

Lin, Changhang, *et al*. (2026) PepGraphormer: an ESM-GAT hybrid deep learning framework for antimicrobial peptide prediction. Journal of Cheminformatics, 18.1: 15.

Malshikare, Hrushikesh, *et al*. (2026) Mechanistic Principles of Antimicrobial Peptides Uncovered by Charge Density Based Machine Learning. Chemical Communications, 2026.

Mera-Banguero, Carlos, *et al*. (2024) AmpClass: an antimicrobial peptide predictor based on supervised machine learning. Anais da Academia Brasileira de Ciências, 2024, 96: e20230756.

Mohkam, Milad; Nezafat, Navid; Ghasemi, Younes. (2024) An in silico approach to de novo design of anti-microbial peptide from inspirited Komodo dragon's original VK6 peptide. International Journal of Data Mining and Bioinformatics, 2024, 28.1: 18-39.

Negovetić, Mario, *et al*. (2024) Efficiently solving the curse of feature-space dimensionality for improved peptide classification. Digital Discovery, 2024, 3.6: 1182-1193.

Pikalyova, Karina, *et al*. (2025) Design of Highly Potent Antibiofilm, Antimicrobial Peptides Using Explainable Artificial Intelligence. Journal of Chemical Information and Modeling, 2025, 66.1: 744-755.

Salimi, Abbas; Lee, Jin Yong. (2026) CoLPAT-AMP: A Transformer-Based framework for Designing novel antimicrobial peptides with property Awareness and partially controllable length. Expert Systems with Applications, 2026, 131999.

Shaon, Md Shazzad Hossain, *et al*. (2024) AMP-RNNpro: a two-stage approach for identification of antimicrobials using probabilistic features. Scientific Reports, 2024, 14.1: 12892.

Shoombuatong, Watshara, *et al*. (2025) Advancing the accuracy of anti-MRSA peptide prediction through integrating multi-source protein language models. Interdisciplinary Sciences: Computational Life Sciences, 2025, 17.3: 716-729.

Singh, Onkar. (2022) Integration of Machine Learning Along With Global Features of the Amino Acid Sequence in Predicting Antimicrobial and Anti-Inflammatory Peptides. 2022. PhD Thesis. National Yang Ming Chiao Tung University.

Teimouri, Hamid; Medvedeva, Angela; Kolomeisky, Anatoly B. (2023) Bacteria-specific feature selection for enhanced antimicrobial peptide activity predictions using machine-learning methods. Journal of Chemical Information and Modeling, 2023, 63.6: 1723-1733.

Yu, Qinze, *et al*. (2026) Uncovering evolutionarily remote and highly potent antimicrobial peptides with protein language models. Nature Biomedical Engineering, 2026, 1-14.

Zahid, Hamza, *et al*. (2026) PeptideNet: An Integrative Deep Learning Framework for Predicting Diverse Bioactive Peptides Using Protein Language Model Embeddings. Journal of Chemical Information and Modeling, 2026.