Briefings in Bioinformatics, 2025, **26(4)**, bbaf343
https://doi.org/10.1093/bib/bbaf343
Problem Solving Protocol

AI-guided discovery and optimization of antimicrobial peptides through species-aware language model

Daehun Bae[1,‡], Minsang Kim[2,‡], Jiwon Seo[2,*], Hojung Nam[1,3,*]

[1]Department of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Buk-gu, Gwangju 61005, Republic of Korea
[2]Department of Chemistry, Gwangju Institute of Science and Technology (GIST), Buk-gu, Gwangju 61005, Republic of Korea
[3]AI Graduate School, Gwangju Institute of Science and Technology (GIST), Buk-gu, Gwangju 61005, Republic of Korea
*Corresponding authors. Jiwon Seo, Department of Chemistry, Gwangju Institute of Science and Technology (GIST), Buk-gu, Gwangju, 61005, Republic of Korea. E-mail: jseo@gist.ac.kr; Hojung Nam, Department of Electrical Engineering and Computer Science, Gwangju Institute of Science and Technology (GIST), Buk-gu, Gwangju, 61005, Republic of Korea. E-mail: hjnam@gist.ac.kr
‡Daehun Bae and Minsang Kim contributed equally.

Abstract

The rise of antibiotic-resistant bacteria drives an urgent need for novel antimicrobial agents. Antimicrobial peptides (AMPs) show promise solutions due to their multiple mechanisms of action and reduced propensity for resistance development. This study introduces LLAMP (Large Language model for AMP activity prediction), a target species-aware AI model that leverages pre-trained language models to predict minimum inhibitory concentration values of AMPs. Using LLAMP, we screened approximately 5.5 million peptide sequences, identifying peptides **13** and **16** as the most selective and most potent candidates, respectively. Analysis of attention values allowed us to pinpoint critical amino acid residues (e.g., Trp, Lys, and Phe). Using the critical amino acids, the sequence of the most selective peptide **13** was engineered to increase amphipathicity through targeted modifications, yielding peptide **13–5** with an overall enhancement in antimicrobial activity but a reduction in selectively. Notably, peptides **13–5** and **16** demonstrated antimicrobial potency and selectivity comparable to the clinically investigated AMP pexiganan. Our work demonstrates the potential of AI to expedite the discovery of peptide-based antibiotics to combat antibiotic resistance.

Keywords: antibiotics; antimicrobial peptides; multi-drug resistant; AI-based screening; protein language model

Introduction

Antibiotic resistance has emerged as one of the most pressing global health challenges, with the World Health Organization warning of a potential 'post-antibiotic era' where common infections could once again become lethal [1]. The rapid spread of multi-drug resistant (MDR) bacteria has outpaced the development of new antibiotics, emphasizing the urgent need for innovative strategies to discover and develop novel antimicrobial agents [2–4]. In recent years, antimicrobial peptides (AMPs) have attracted considerable attention as potential alternatives to traditional antibiotics [5–7]. Characterized by their multiple mechanisms of action, such as disrupting cell membranes and targeting intracellular nucleic acids and proteins, and their lower likelihood of promoting resistance development, AMPs have garnered significant attention as promising candidates for next-generation antibiotics. However, despite their considerable potential, the discovery of novel AMP lead sequences often requires the high expenses and prolonged timelines inherent in traditional structure–activity relationship (SAR)-based experiments.

Artificial intelligence (AI) has emerged as a transformative force in the face of these challenges, altering the field of antibiotic discovery. The adoption of AI technologies in the drug discovery pipeline has demonstrated the potential to accelerate both the identification and optimization of innovative therapeutic agents [8–10]. In the context of AMP discovery, AI can be employed to predict the antimicrobial activity of peptides, enabling the rapid screening of vast libraries and the identification of promising hit peptides for further experimental validation [11–16]. AI models for AMP discovery are based on the physicochemical properties of peptide molecules as inputs and exploit various architectures, including deep neural networks, convolutional neural networks (CNN) [17], long short-term memory (LSTM) networks [18], and bidirectional LSTM (Bi-LSTM) networks [19]. Moreover, the recent emergence of language models based on transformer architecture [20] is increasingly being integrated into the AMP field. Protein language models have gained significant traction in AMP research. Models such as ProtTrans, ESM-1b, and ESM-2 have been widely adopted [21–23]. These models excel at processing peptide sequences as input, autonomously learning, and leveraging relevant features. This capability has proven instrumental in facilitating the discovery of novel AMPs, as demonstrated in several recent studies [24–26].

Despite their numerous strengths, previous studies have encountered significant limitations, particularly in their methodology for classifying peptide sequences as either AMPs or non-AMPs. Additionally, a notable deficiency in many of these studies

Received: March 19, 2025. **Revised:** June 5, 2025. **Accepted:** June 22, 2025

is the lack of consideration for the target bacterial species within their predictive models. To address these challenges, we introduce LLAMP (Large Language model for AMP activity prediction), a species-aware activity predictor. LLAMP represents an innovative AI model that employs advanced deep learning techniques, leveraging the power of pre-trained language models. Based on ESM-2, a protein language model utilizing transformer architecture, LLAMP is designed to predict the minimum inhibitory concentration (MIC) values of peptide sequences against different bacterial species. Our approach involved several key steps (Fig. 1a, b). First, we collected 1.7 million extensive peptide sequence data. Second, we applied ESM-2 for masked language modeling (MLM), resulting in a peptide-tuned version of ESM-2. Next, we encoded genomic information of various bacterial species to represent the target species. Finally, we fine-tuned the peptide-tuned ESM-2 model, integrating the genomic information to create LLAMP. The LLAMP model is capable of predicting MIC values for AMPs based on both the peptide sequence and the target bacterial species. By providing species-aware MIC predictions, LLAMP aims to improve the efficiency of AMP design and screening processes. This is particularly valuable in the fight against newly emerging MDR bacteria, where rapid and targeted antibiotic development is crucial.

The rationale for employing AI in AMP discovery stems from the need to rapidly explore the vast chemical space of potential peptide sequences. Traditional SAR studies, limited by throughput and cost, struggle to keep pace with the rapid emergence of MDR pathogens. LLAMP addresses these limitations by enabling the screening of approximately 5.5 million peptide sequences, drastically reducing the time and resources required to identify promising AMPs (Fig. 1c). From the 16 peptide sequences suggested by LLAMP, we selected 7 sequences based on physicochemical properties (e.g., hydrophobicity, and amphipathicity) and predicted secondary structure (e.g., helical or non-helical) (Fig. 1c). This selection process was benchmarked against well-known AMPs (i.e., pexiganan, indolicidin, and LL-37). The selected sequences were synthesized and evaluated for their antimicrobial and hemolytic activities. Moreover, the attention mechanisms within LLAMP provide valuable insights into the significance of specific amino acid residues. This information guides the optimization of peptide sequences, further enhancing their antimicrobial activity. By combining AI-driven screening with targeted experimental validation, our approach offers a powerful new tool in the ongoing battle against antibiotic resistance.

Results

LLAMP is a well-trained peptide language model

In this study, we introduce LLAMP, a cutting-edge AI model designed for species-aware antibiotic activity prediction. LLAMP leverages advanced machine learning techniques, including the use of pre-trained language models. At its core is ESM-2 [23], a protein language model built on the transformer architecture, which is utilized to predict the MIC values for peptide sequences targeting specific bacterial species.

The development process began with the collection of peptide sequence data (please see Datasets in Materials and Methods for more details), on which ESM-2 was applied for masked language modeling (MLM). This process involved masking a portion of the amino acids in each peptide sequence and training the model to predict the masked tokens, leading to the generation of a peptide-optimized version of ESM-2 (peptide-tuned ESM-2).

To verify the proficiency of the peptide-tuned ESM-2 model in interpreting peptide language, we assessed its accuracy in MLM prediction and pseudo-perplexity using a validation set of 69,242 peptide sequences. The accuracy of MLM prediction was evaluated by comparing the model's top-k predicted amino acids at each masked position with the actual amino acids (Fig. 2a). The peptide-tuned ESM-2 model demonstrated higher accuracy compared to the original ESM-2 model in predicting the masked amino acids, indicating its ability to capture the underlying patterns and dependencies in peptide sequences.

Pseudo-perplexity (PPPL), a measure used to evaluate the uncertainty of a language model about a given sequence of text, was also employed to assess the model's performance [27]. As shown in Fig. 2b, the pseudo-perplexity of the peptide-tuned ESM-2 model decreased significantly during the training process, reaching its minimum value of 14.0, compared to the initial value of 18.9 for the original ESM-2 model. This reduction in pseudo-perplexity indicates a substantial improvement in the model's understanding of peptide languages and ability to predict masked residues of peptide sequences after pre-training.

To further investigate the efficacy of pre-training, we analyzed the alterations in amino acid token embeddings, both pre- and post-training, utilizing UMAP visualization [28]. The results, depicted in Fig. 2c, exhibit a pronounced shift in the embeddings from the baseline ESM-2 model to the peptide-tuned ESM-2. This transition suggests a refined adaptation to peptide-specific languages. Remarkably, the amino acid embeddings demonstrate a tendency to cluster by their respective properties, providing compelling evidence that the peptide pre-training has significantly enhanced the model's capacity to comprehend and interpret peptide language.

The peptide-tuned ESM-2 model's enhanced performance in MLM prediction accuracy and reduced pseudo-perplexity demonstrates its effectiveness in capturing the intricate patterns and dependencies within peptide sequences. This optimized model serves as a strong foundation for the subsequent development of the LLAMP framework, enabling more accurate predictions of antimicrobial peptide activity against specific bacterial species.

LLAMP outperforms in the prediction of species-aware AMP activity

To confirm the predictive performance of the LLAMP model at the species level, we conducted a comprehensive evaluation of its species-aware performance using the fine-tuning data set collected from DBAASP v3 (please see Datasets in Materials and Methods for more details, Supplementary Figs. 1 and 2, Supplementary Section 1). The performance of LLAMP was assessed on 35 target species, and the results are presented in Supplementary Table 1. Notably, 54% of the species exhibited Pearson correlation coefficient (PCC) greater than 0.7, indicating a strong predictive performance. However, certain species, such as *Listeria innocua*, showed limited predictive capability, likely due to insufficient training data. This highlights the importance of having a diverse and representative dataset for training the model to ensure robust performance across a wide range of species.

The resulting model, LLAMP, was then compared against five machine learning-based models, two deep learning-based models, and two BERT [29]-based baseline models using the preprocessed test dataset (Table 1). The details of the baseline models can be found in Supplementary Fig. 3 and Supplementary Section 2. The average prediction performance of LLAMP and the baseline models for 35 target species was evaluated using a test dataset.

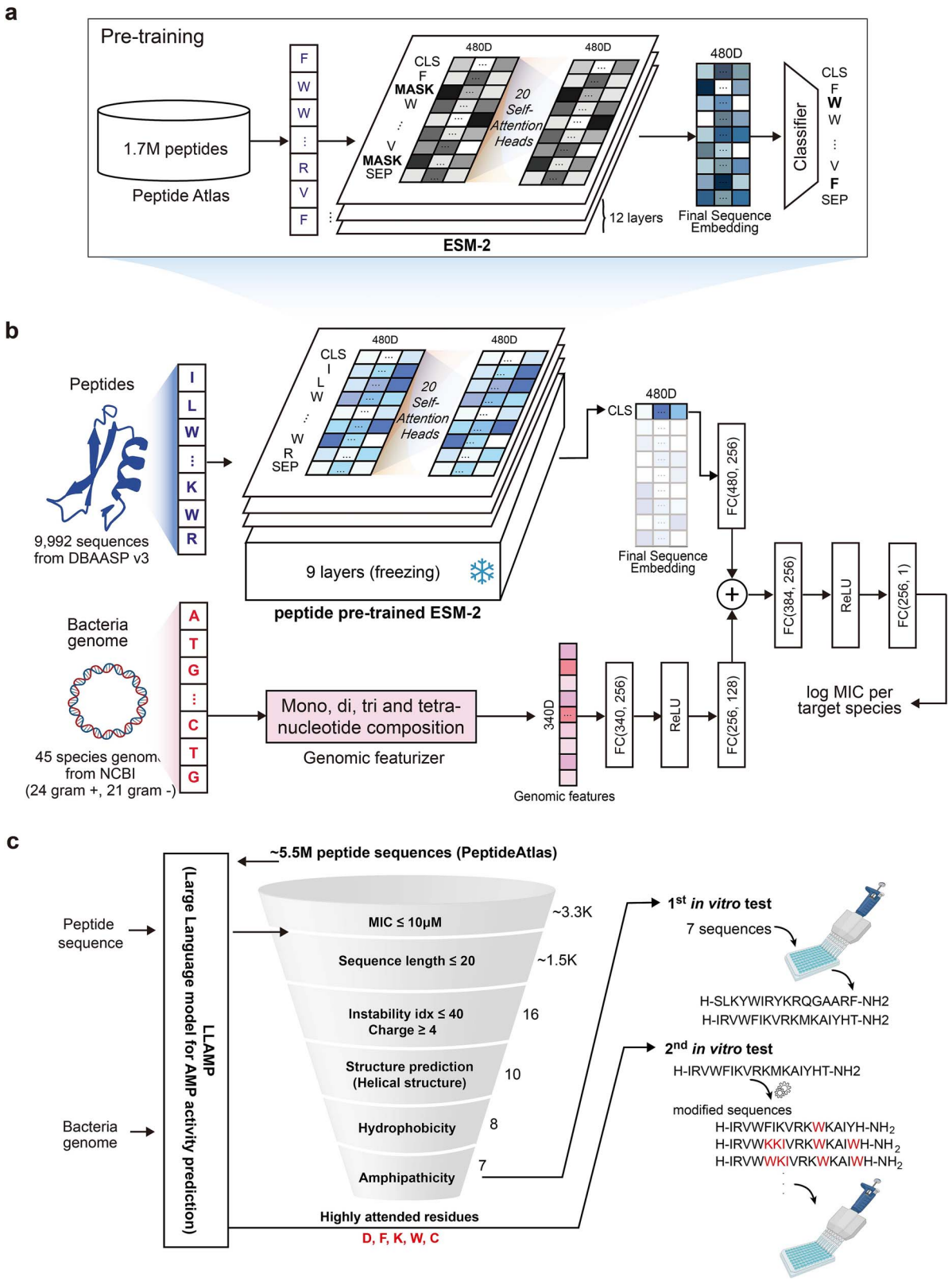

Figure 1. AI-based AMP discovery. (a) Pre-training process of the LLAMP model. In the pre-training process, the ESM-2 protein language model is trained with ~1.7 M peptide sequence so that the model can learn sequence patterns of peptides. (b) In the fine-tuning process, the LLAMP model learns species-aware activities of peptides and predicts MIC values. (c) AI-based screening processes. In the screening processes, the LLAMP model and peptide properties are used. Alt text: Workflow illustrating AI-based antimicrobial peptide discovery: Pre-training a language model on peptide sequences, fine-tuning for MIC prediction, and screening peptides using model outputs and properties.

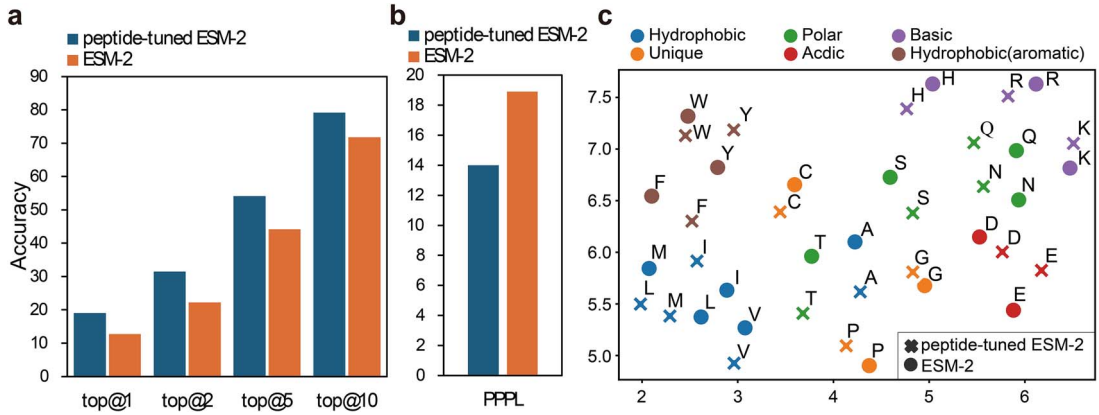

Figure 2. Evaluation of the peptide-tuned ESM-2 model and visualization of amino acid embeddings. (a) the accuracy of masked language modeling (MLM) prediction results for the top-k most probable amino acids at each masked position. (b) Pseudo-perplexity (PPPL) plot comparing the performance of the peptide-tuned ESM-2 model with the original ESM-2 model on a validation set of 69,242 peptide sequences. Lower pseudo-perplexity indicates better performance in capturing the peptide language. (c) UMAP visualization of amino acid token embeddings derived from both the original ESM-2 and peptide-tuned ESM-2 models. The peptide-tuned ESM-2 embeddings show distinct clustering based on amino acid properties, demonstrating the model's improved understanding of peptide language. Alt text: Performance comparison of original and peptide-tuned ESM-2 models, including masked language modeling accuracy, pseudo-perplexity plots, and UMAP visualizations showing improved clustering of amino acid embeddings by biochemical properties.

Table 1. Dataset statistics for training, validation, and testing in peptide MLM tuning and regression fine-tuning

Dataset	Origin	Split type	# Peptides	# Species	# Entities
Pre-training	PeptideAtlas [33]	Train	1,661,800	-	-
		Validation	69,242	-	-
		Total	1,731,042	-	-
Fine-tuning	DBAASP v3 [41]	Train	7344	35	29,394
		Validation	918	35	3683
		Test	919	35	3741
		Both-unseen	811	10	982
		Total	9992	45	37,800

As presented in Table 2, LLAMP demonstrated superior performance across all metrics, including PCC of 0.735, coefficient of determination (R^2) of 0.536, mean absolute error (MAE) of 0.379, mean squared error (MSE) of 0.272, and root mean squared error (RMSE) of 0.521. These results highlight the overall effectiveness of LLAMP in accurately predicting MIC values for AMPs based on their peptide sequences, outperforming the baseline models.

To further investigate the model's performance under more challenging conditions, we employed a both-unseen dataset, which consists of peptide sequences and target species that were not encountered during the training process. This dataset allows for a rigorous assessment of the model's generalizability and ability to handle novel data. The species-specific performance metrics for the both-unseen set are detailed in Supplementary Table 2. Remarkably, 50% of these species exhibited a PCC score exceeding 0.7, reinforcing our confidence in the model's capability to accurately predict AMP activity against novel species that were not part of the training data. These outcomes demonstrate the robustness of LLAMP across a variety of species and highlight its potential utility in identifying active AMPs effective against newly emerged or antibiotic-resistant species. Also, as shown in Table 2, LLAMP exhibited superior performance compared to the baseline models on this both-unseen dataset, with a PCC of 0.668, R^2 of 0.418, MAE of 0.428, MSE of 0.342, and RMSE of 0.585. Regarding run time complexity, BERT-based models, including LLAMP, were more computationally intensive but delivered significantly higher accuracy and species-aware capabilities essential for targeted AMP development (Supplementary Section 3, Supplementary Table 3).

To compare state-of-the-art (SOTA) deep learning models for AMP identification, we trained and tested four models (AmPEPpy [30], AMPScannerV2 [15], Cao et al. [31], and Diff-AMP [32]) using derived classification dataset from LLAMP dataset. A classification dataset was constructed from the LLAMP dataset by assigning activity labels based on MIC thresholds. Peptide-bacterial species pairs with MIC values below 25 μM were classified as active, whereas those with MIC values exceeding 100 μM were classified as inactive. Further details can be found in Supplementary Section 4 and Supplementary Table 4. As shown in Table 3, LLAMP demonstrated superior performance, achieving an accuracy (ACC) 90.53, outperforming AmPEPpy, AMPScannerV2, Cao et al., and Diff-AMP. LLAMP achieved the best performance in terms of ACC, specificity (SPE), Matthews correlation coefficient (MCC) and F1-Score. Although AmPEPpy shows better performance in terms of sensitivity (SEN), this was due to its tendency to bias most of its predictions toward the Active class. LLAMP demonstrated the most balanced and highest performance in bacterial genome-based predictions. In real-world applications, it is essential to consider the target bacterial strains. Therefore, LLAMP is more useful in early screening steps.

Lastly, to further validate LLAMP's performance, we compared it to previous models that predict MICs for specific target species[13–14], using datasets sourced from their respective works. As shown in Table 4, when training, validating, and testing

Table 2. Prediction performances on the test and both-unseen sets

Set	Type	Method	PCC	R^2	MAE	MSE	RMSE
Test set	Machine learning-based architectures	RF [48] + CTD [49]	0.709	0.491	0.423	0.298	0.546
		XGB [50] + CTD	0.651	0.407	0.441	0.347	0.589
		SVM [51] + CTD	0.627	0.390	0.454	0.357	0.600
		RF + ESM-2	0.680	0.443	0.448	0.326	0.571
		XGB + ESM-2	0.644	0.407	0.452	0.347	0.589
	Deep learning-based architectures	CTD + FC	0.581	0.317	0.495	0.400	0.633
		Embedding + CNN	0.680	0.451	0.420	0.322	0.567
	BERT-based architecture	protBERT-BFD [21]	0.644	0.390	0.443	0.358	0.598
		ESM-2	0.723	0.523	0.388	0.280	0.529
		LLAMP (proposed)	**0.735**	**0.536**	**0.379**	**0.272**	**0.521**
Both-unseen set	Machine learning-based architectures	RF + CTD	0.640	0.402	0.458	0.352	0.593
		XGB + CTD	0.608	0.335	0.475	0.391	0.626
		SVM + CTD	0.480	0.180	0.539	0.482	0.695
		RF + ESM-2	0.650	0.415	0.443	0.344	0.587
		XGB + ESM-2	0.577	0.273	0.487	0.428	0.654
	Deep learning-based architectures	CTD + FC	0.541	0.243	0.511	0.445	0.667
		Embedding + CNN	0.604	0.300	0.487	0.412	0.642
	BERT-based architecture	protBERT-BFD	0.646	0.363	0.457	0.375	0.612
		ESM-2	0.618	0.337	0.469	0.390	0.625
		LLAMP (proposed)	**0.668**	**0.418**	**0.428**	**0.342**	**0.585**

Table 3. Evaluation and comparison of multiple AMP recognition models based on performance (%)

Method	ACC	SEN	SPE	MCC	F1-Score
AmPEPpy	87.45	**96.26**	49.26	54.13	92.57
AMPScannerV2	87.31	95.32	49.67	51.61	92.53
Cao *et al.*	85.86	93.66	52.03	50.07	91.50
Diff-AMP	83.16	89.45	55.90	45.07	89.62
LLAMP	**90.53**	95.79	**67.71**	**67.40**	**94.27**

using the ESKAPEE dataset, LLAMP demonstrated similar or better performance on the test set compared to the other models. As shown in Supplementary Table 5, ESKAPEE-MICpred requires a larger number of parameters for inference and is more complex due to the ensemble of four models. However, LLAMP, despite using fewer parameters, achieves better performance and offers greater efficiency during inference. It is important to note that while both LLAMP and ESKAPEE-MICpred can predict species-aware AMP activity, ESKAPEE-MICpred is only able to support seven species and lacks the generalization ability to predict activity for new species. In contrast, LLAMP, trained on a broader range of species, exhibited a stronger predictive ability for new species, highlighting its greater versatility and generalizability. Although LLAMP has a large number of model parameters compared to GRAMPA(CNN), LLAMP showed a notable improvement in performance on the GRAMPA-*E.coli* set, with a PCC increase from 0.770 to 0.811 and a decrease in RMSE from 0.501 to 0.454. This improvement highlights the effectiveness of LLAMP in accurately predicting AMP activity against specific bacterial species, even when compared to models designed explicitly for those species.

The ablation study unveils the critical roles of pre-training and fine-tuning

To identify the key factors contributing to the model's enhanced performance, we conducted an ablation study by testing the model under various configurations. These cases included (i) without peptide pre-training, (ii) with all transformer encoder layers frozen (all freeze), (iii) without the entire genomic branch, and (iv) with a reduced number of species in training (Supplementary Fig. 4).

In the 'without peptide pre-training' scenario, we used the original ESM-2 model directly for fine-tuning, without any peptide-specific pre-training. The 'all freeze' scenario involved freezing the transformer encoder layers during fine-tuning, while the 'reduced species' scenario entailed training with only eight species. We evaluated the results of these ablation experiments on the test set (cases 1 and 2; Supplementary Fig. 5a and Supplementary Table 6) and the both-unseen set (cases 1, 2, and 3; Supplementary Fig. 5b and Supplementary Table 6).

The 'all freeze' scenario led to a substantial decrease in performance on both sets, highlighting the crucial role of fine-tuning the transformer encoder layers. Fine-tuning these layers appears to stabilize the model and facilitate its specialization in regression tasks, enabling it to capture the nuances and complexities of the antimicrobial activity prediction problem. Also, the 'without genomic branch' case shows a significant performance drop, confirming the essential role of species-specific genomic information in accurate MIC prediction. Conversely, the absence of peptide pre-training resulted in a slight performance drop, suggesting that peptide pre-training plays a valuable role in enhancing the model's understanding of peptide languages and their inherent properties. Lastly, our analysis revealed that decreasing the number of species used in training from 35 to 8 led to worse

Table 4. Performance of our model and previous species-aware studies

Method	Dataset	PCC	R^2	MAE	MSE	RMSE
ESKAPEE-MICpred	ESKAPEE [14]	0.802	0.640	0.442	0.376	0.613
LLAMP		**0.805**	**0.648**	**0.422**	**0.365**	**0.604**
GRAMPA (CNN)	GRAMPA-*E. coli* [13]	0.770	-	-	-	0.501
LLAMP		**0.811**	0.656	0.335	0.206	**0.454**

Table 5. Sequences and properties of antimicrobial peptides (AMPs) predicted by large language model for AMP prediction (LLAMP)

#	Sequence	CTLR[a]	UPLC elution (%CH_3CN)
2	H-AKQGFVIYRVRVRRGNRK-NH_2	0.39	37.4
4	H-FKRVWNYFQRVLVKKYA-NH_2	0.29	45.1
10	H-FKRVWNYFQRVLVKK-NH_2	0.33	44.4
11	H-FKRVWNYFQRVLVKKY-NH_2	0.31	45.3
13	H-IRVWFIKVRKMKAIYHT-NH_2	0.29	43.5
14	H-FSLKYWIRYKRQGAARF-NH_2	0.29	43.7
16	H-SLKYWIRYKRQGAARF-NH_2	0.31	43.7
Pexiganan	H-GIGKFLKKAKKFGKAFVKILKK-NH_2	0.41	46.9

[a]Charge to length ratio (CTLR) was proportion between the number of cationic residues and the whole number of residues in a peptide. Histidine was not considered as a cationic residue.

performance on the bot-unseen set (Supplementary Fig. 5b). This finding highlights the importance of a diverse training set in improving the model's generalization ability. By exposing the model to a wider range of species during training, it can learn to capture the variability and complexity of antimicrobial activity across different bacterial species, resulting in more robust and accurate predictions.

AI-based screening discovers AMP hit sequences

To discover novel AMP candidates, we performed a comprehensive screening of approximately 5.5 million peptides from the PeptideAtlas database [33] (Fig. 1c). Our screening process involved a multi-step filtering approach (detailed in the PeptideAtlas screening section of Materials and Methods). We began by selecting peptides with predicted MIC values below 10 μM against more than 90% of either gram-positive or gram-negative bacterial species among the 45 species. We further refined our selection by choosing peptides with an instability index of 40 or less, which is indicative of protein stability according to Guruprasad et al. [34]. Additionally, we kept sequences with a net cationic charge of 4 or greater (Supplementary Fig. 6 and Supplementary Data 1, 2).

Next, we further sorted out based on the physicochemical properties, such as hydrophobicity [35] and amphipathicity [36] (Table 5). Finally, we selected seven peptides **2**, **4**, **10**, **11**, **13, 14,** and **16** based on their amphipathic property induced by α-helix structure. The other peptides in the 16 suggested peptides were not synthesized because they showed similar sequences compared to the selected seven peptides or they exhibited suboptimal properties (e.g., too hydrophilic or lack of cationic charges) or structures (e.g., random coil), compared to the benchmark AMPs (e.g., pexiganan, indolicidin, and LL-37).

In Table 6, the antimicrobial activities of seven AI-predicted sequences against gram-positive and gram-negative pathogenic strains are summarized. Natural AMPs (e.g., melittin and buforin-II) and clinically investigated AMPs (e.g., pexiganan and omiganan) were used as positive controls in the assay. Among the seven AMPs, peptide **16** demonstrated the most potent activity for gram-negative *Escherichia coli* and gram-positive *Staphylococcus aureus*, while exhibiting low hemolytic activity. Compared to pexiganan, peptide **16** showed better or similar antimicrobial activity against gram-negative *Pseudomonas aeruginosa, Salmonella typhimurium, Enterobacter spp.* and gram-positive *S. aureus*, while exhibiting approximately half the hemolytic toxicity (i.e., H_{max} value). Peptide **13** demonstrated the lowest hemolytic activity (e.g., H_{max} value), comparable to that of omiganan, while maintaining relatively potent antimicrobial activity. Peptide **13** showed similarly potent antimicrobial activity compared to pexiganan, against *Acinetobacter baumannii*, *Enterobacter spp.*, and *S. typhimurium*. Importantly, some of these pathogens belong to the ESKAPE group (*Enterococcus faecium*, *S. aureus*, *Klebsiella pneumoniae*, *A. baumannii*, *P. aeruginosa*, and *E. species*), which are currently recognized as a major threat to public health. Peptides **2**, **11**, and **14** showed moderate antimicrobial activity, but peptide **11** exhibited severe hemolysis. Peptides **4** and **11** showed low selectivity between bacterial and mammalian cells likely due to the hydrophobic property of the two peptides (estimated by the UPLC elution (%CH_3CN), Table 5). Peptide **10** did not show antimicrobial activity at concentration lower than 200 μM against gram-negative *E. coli* and gram-positive *S. aureus*.

Increased amphipathicity leads to enhanced antimicrobial activity

Based on the sequence of most selective peptide **13**, we attempted to find a new sequence that has increased antimicrobial potency by site specific mutation of the original sequence. To design the derivatives of peptide **13**, we analyzed the attention matrix from the LLAMP model during the prediction of highly active AMP sequences. Our analysis revealed that specific residues, such as Lys, Phe, Asp, Trp, and Cys, showed high attention in potent AMPs (hypergeometric test, $P < 0.05$, details described in Fig. 3). We then generated sequence logos of 5-amino acid segments centered around each highly attended residue, allowing us to identify frequently appearing motifs (Supplementary Fig. 7). These highly attended residues and their associated motifs served as

Table 6. Antimicrobial activities against pathogenic strains and hemolytic activities of seven AI-predicted AMPs.

Compounds	MIC[a] (μM)										HD_{10}/HD_{50}[b] (μM)	H_{max}[c] (200 μM)	Selectivity index[d]
	Escherichia coli ATCC 25922	*Acinetobacter baumannii* ATCC 19606	*Klebsiella pneumoniae* ATCC 700603	*Pseudomonas aeruginosa* ATCC 27853	*Salmonella typhimurium* SL 1344	*Enterobacter spp.* KCTC 2625	*Staphylococcus aureus* ATCC 25923	*Bacillus subtilis* KCTC 3068	*Enterococcus faecalis* KCTC 3511	*Listeria monocytogenes* KCTC 13064	*Enterococcus faecium* KCTC 13225	*Staphylococcus epidermidis* KCTC 1917			
melittin[e]	12.5	<3.1	12.5	12.5	6.3	<3.1	≤1.6	<3.1	3.1	nd[f]	nd[f]	nd[f]	<3.1/4.0	105.7 ± 17.6	<0.2
buforin-II[e]	12.5	50	>100	100	>100	50	12.5	50	50	nd[f]	nd[f]	nd[f]	nd[f]	nd[f]	nd[f]
omiganan[e]	12.5	12.5	25	12.5	25	<3.1	12.5	6.3	25	nd[f]	nd[f]	nd[f]	>100/>100	0	>8.0
pexiganan	6.3	<3.1	<3.1	25	6.3	<3.1	25	6.3	6.3	3.1	6.3	3.1	>100/>100	13.5 ± 1.8	>15.9
2	25	6.3	25	>100	12.5	3.1	100	3.1	>100	>100	>100	100	>100/>100	0.9 ± 0.2	>4.0
4	50	3.1	12.5	>100	6.3	>100	200	<3.1	>100	50	25	50	48.9/>100	49.9 ± 1.4	1.0
10	200	>100	25	100	25	>100	>200	3.1	50	50	6.3	>100	>100/>100	2.5 ± 0.4	>0.5
11	50	<3.1	12.5	>100	6.3	<3.1	100	<3.1	>100	50	25	25	39.2/>100	62.7 ± 0.2	0.8
13	25	<3.1	25	>100	3.1	<3.1	50	3.1	100	50	25	25	>100/>100	0.6 ± 0.4	>4.0
14	25	6.3	>100	>100	25	<3.1	100	6.3	>100	100	25	100	>100/>100	4.2 ± 0.1	>4.0
16	12.5	6.3	12.5	6.3	6.3	<3.1	25	6.3	12.5	25	25	25	>100/>100	5.7 ± 1.4	>8.0

[a]These concentrations represent mean values in triplicate. [b]HC_{10} and HC_{50} are the concentrations of compounds causing 10% and 50% hemolysis in rat erythrocytes, respectively. These concentrations represent mean values in triplicate. [c]H_{max} is the percentage (%) of hemolysis at the highest concentration tested (200 μM). [d]The selectivity index was calculated as HC_{10} divided by the minimum inhibitory concentration (MIC) values in *E. coli* ATCC 25922. [e]Literature values of melittin, buforin-II, and omiganan were obtained from ref [52]. [f]Not determined.

references for further optimization and design of the second generation library.

Using the highly attended amino acids from the attention matrix analysis (Fig. 3), we mutated amino acid in peptide **13**, generating six new derivative sequences (Supplementary Table 7). The direction of mutation was to increase amphipathicity and hydrophobic moment, crucial structural and physico-chemical properties for bacterial membrane disruption (Fig. 4). Details of the derivative sequence generation process are shown in Section 5 of **Supplementary Information**.

In Table 7, the antimicrobial activities of peptide **13** derivatives against gram-positive and gram-negative pathogenic strains are summarized. Compared to peptide **13**, peptides **13–1** and **13–2** were modified in amino acids at the hydrophobic surface and demonstrated reduced antimicrobial activities and increased hemolytic toxicities. Peptides **13–3** and **13–4** adopted substitutions and truncation to reinforce hydrophobic moment and amphipathic property, but they showed decreased antimicrobial activities with reduced hemolytic toxicities. To improve biological activities, tryptophan residues were incorporated into the hydrophobic surface of peptide **13–3**, resulting in the design of peptide **13–5** with enhanced amphipathic properties. Notably, this modification improved its antibacterial activity against *E. coli* and *P. aeruginosa*. Peptide **13–6**, which lysine was substituted to tryptophan, maintained its activity against gram-negative bacteria and demonstrated enhanced activity for *S. aureus*. While **13–5** showed retained hemolytic toxicity compared to the parent peptide, **13–6** exhibited increased hemolytic toxicity.

Generally, **13–5** maintained moderate to strong antimicrobial activity across a diverse range of both gram-negative and gram-positive bacteria, particularly with similar MIC values compared to pexiganan against *A. baumannii*, *Enterobacter spp.*, *S. aureus* and *Enterococcus faecalis*. Moreover, AI-derived peptides **16** and **13–5** demonstrated antimicrobial potency and selectivity comparable to the clinically investigated AMP pexiganan. These results highlight the potential of our AI-guided approach to designing effective and potentially safer antimicrobial peptides, with peptides **13**, **13–5**, and **16** emerging as lead candidates that warrant further investigation and optimization.

Conformation and mechanism study

To evaluate the relationship between peptide conformation and activity, CD spectra of peptides were measured in three different conditions (Fig. 5). As previously reported [37], pexiganan exhibited random coil structures in both Tris–HCl aqueous buffer and erythrocyte membrane mimic conditions as shown in Fig. 5j. On the other hand, pexiganan formed a helical conformation under *E. coli* membrane mimic conditions, as indicated by its CD signature. This type of conformational behavior explains the enhanced selectivity of pexiganan for bacterial cells over eukaryotic cells. Interestingly, peptide **13–5** showed a similar conformational transition under the three different conditions, which accounts for its high selectivity. Unlike **13–5**, peptide **16** did not indicate any evidence of helix formation in the aqueous buffer or membrane-mimicking conditions. Supplementary Section 6 and Section 7 contains additional information.

To investigate the mechanism of the peptides, nitrocefin assay and PI uptake assay were performed (Fig. 6). These two assays can monitor the disruption of the outer membrane and the cytoplasmic membrane, respectively. For both nitrocefin assay and PI uptake assay, melittin was used as a positive control, and buforin-II was used as a negative control. Peptide **13** with the

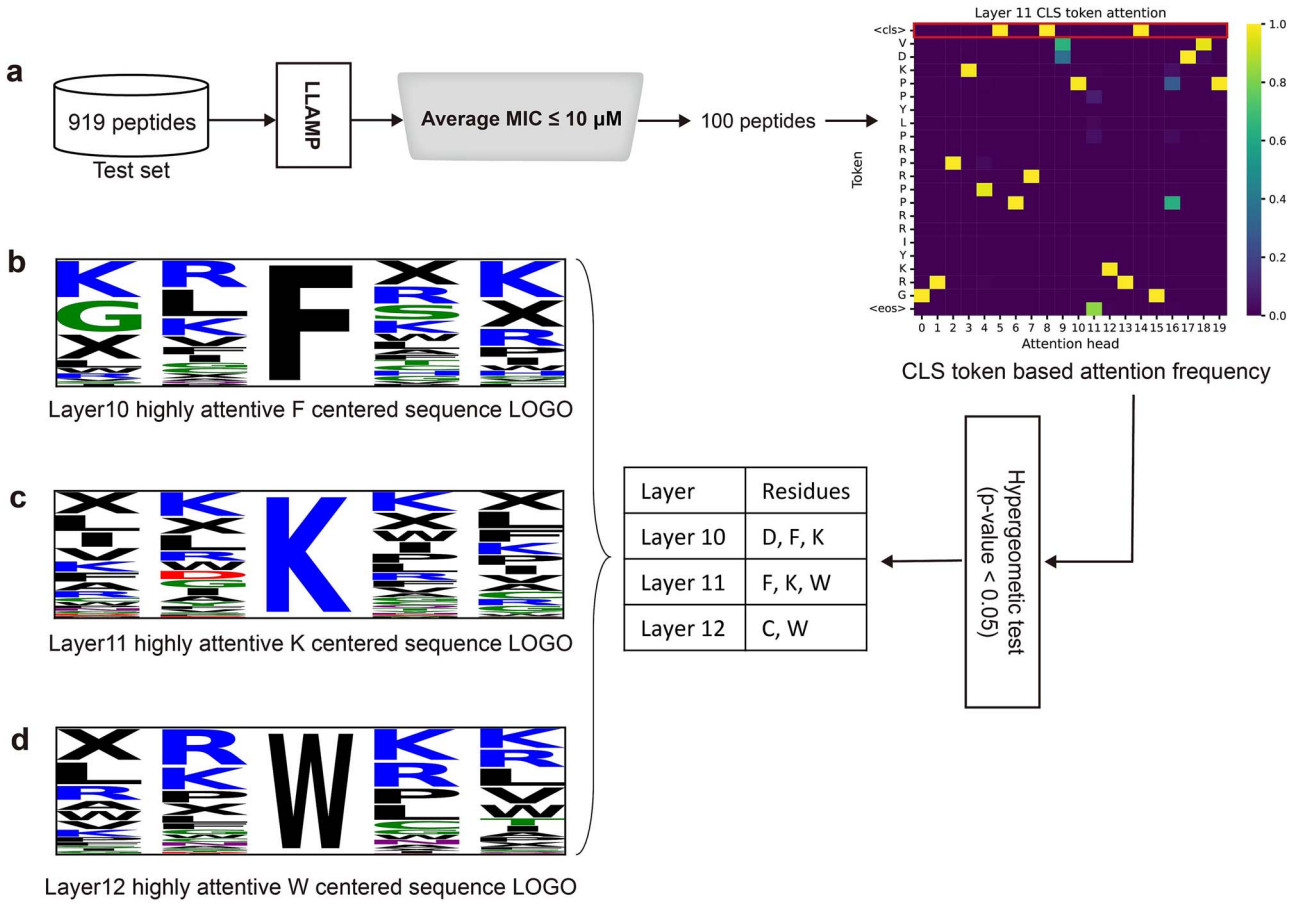

Figure 3. Attention analysis based AMP sequence optimization. (a) Whole attention analysis pipeline. In attention analysis, LLAMP CLS token attention is used with 100 active peptides so that the residue-wise frequencies are calculated and important residues are extracted. (b), (c) and (d) are highly attentive F (phenylalanine), K (lysine) and W (tryptophan) centered sequence LOGO for each layers 10, 11, and 12. X residue means terminal padding residue. Alt text: Diagram of the attention analysis process for AMP sequence optimization, showing identification of key residues and sequence logo plots highlighting frequent phenylalanine, lysine, and tryptophan positions across model layers.

highest GRAVY score [35], a parameter of hydrophobicity, showed moderate membrane disruptive effects on both the outer and inner bacterial membranes. Similarly, peptide **16**, identified as the most potent AI-predicted AMP, disrupted bacterial membranes, though its effect on the inner membrane was noticeably weaker. Peptide **2**, the most hydrophilic peptide in the library, displayed minimal membrane-disruptive activity on both the outer and inner membranes. In contrast, peptide **14** exhibited strong outer membrane disruption and moderate inner membrane disruption. The modified peptide **13–5** and **13–6**, based on the peptide **13**, effectively disrupted the outer membrane compared to melittin, and indicated a greater degree of inner membrane disruption than their parent peptide **13**. These results indicate that increased hydrophobicity and amphipathicity by the sequence modifications in **13–5** and **13–6**, led to greater bacterial membrane disruption compared to **13**. Additional details are provided in Supplementary Section 8 and Section 9.

Discussion

Our study demonstrates the successful development and application of LLAMP, a species-aware AI model that advances the field of antimicrobial peptide discovery in several key aspects. First, LLAMP's superior performance compared to existing prediction models (PCC of 0.735 and R^2 of 0.536 on the test set) validates our approach of combining peptide pre-training with species-specific genomic features. Particularly noteworthy is LLAMP's ability to generalize to previously unseen bacterial species, achieving a PCC of 0.668 on the both-unseen dataset. This capability is crucial for addressing emerging pathogens and MDR bacteria.

The discovery of peptides **13** and **16** through AI-guided screening exemplifies LLAMP's practical utility. Peptide **16** demonstrated remarkable potency against both gram-negative and gram-positive bacteria, with MIC values comparable to or better than pexiganan for several ESKAPE pathogens, while maintaining low hemolytic activity. The selective activity of peptide **13** against bacterial cells over erythrocytes, comparable to omiganan, highlights the model's ability to identify candidates with desirable therapeutic properties. Due to the high cost of peptide synthesis, the number of peptides actually synthesized was limited to seven. However, in the predictive model currently under development, the experimental dataset is being expanded by focusing on more clinically relevant pathogens, such as *P. aeruginosa*, *A. baumannii*, and multidrug resistant strains.

LLAMP's attention analysis provided valuable mechanistic insights, revealing the importance of specific amino acids like Trp, Lys, and Phe in antimicrobial activity. This information guided the rational design of second-generation peptides, leading to peptide **13–5**, which showed enhanced antimicrobial activity while retaining favorable selectivity. The successful optimization of peptide **13** demonstrates how AI-derived insights can inform

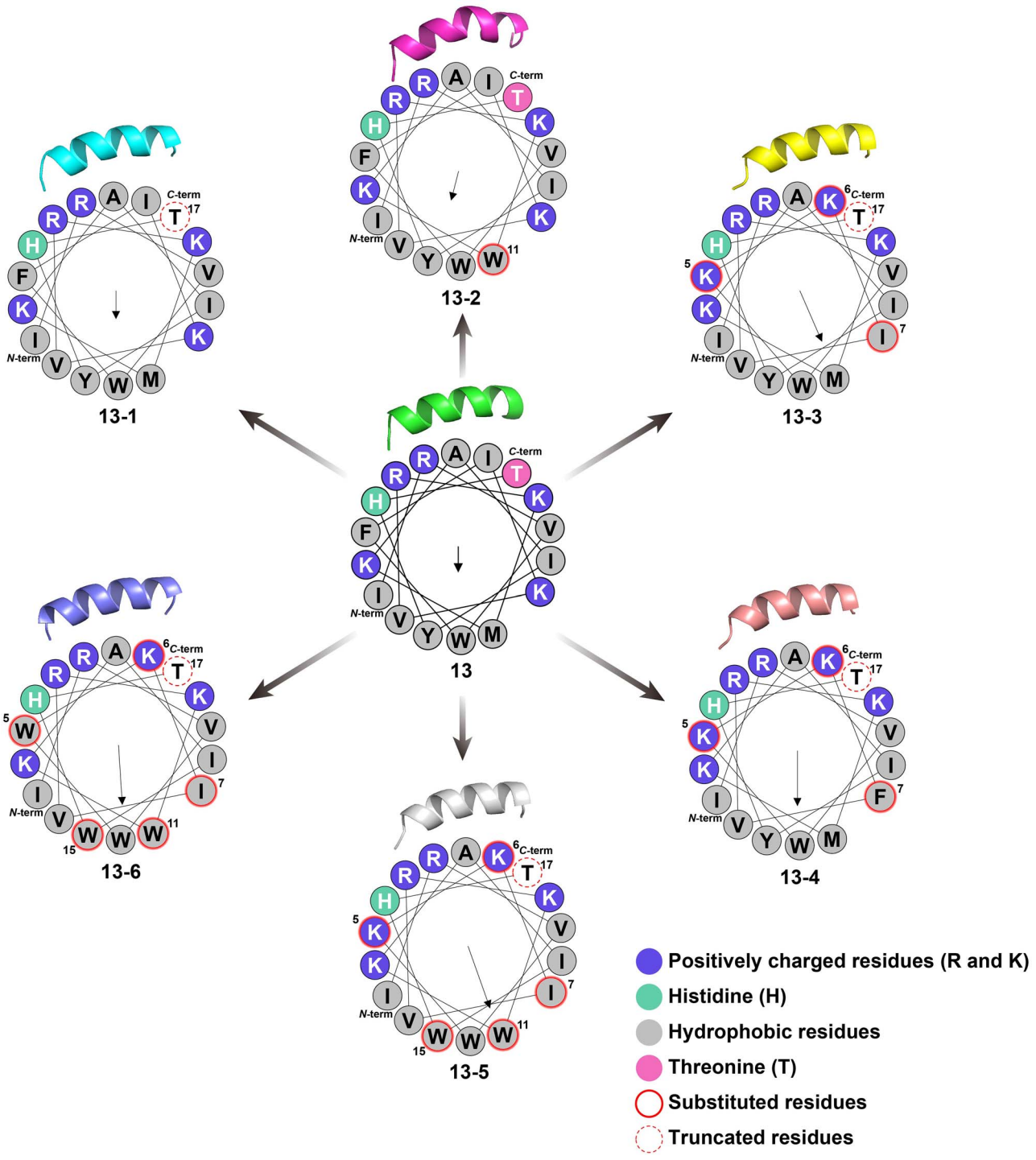

Figure 4. Helical wheel diagram and predicted 3D structure of peptide 13 and its derivatives generated using HeliQuest [47]. Arrows indicate the direction and magnitude of the hydrophobic moment. Residues labeled with 'N-term' or 'C-term' denote the N-terminus or C-terminus of the peptide, respectively. Alt text: Structural diagrams comparing peptide 13 derivatives to the parent peptide: Helical wheels show residue types (charged/hydrophobic) and modifications (substitutions/truncations), while 3D models display hydrophobic moment vectors and termini positions.

peptide engineering strategies. The conformational studies of peptides **13–5** and **16** revealed distinct mechanisms of action. Peptide **13–5**'s conformational behavior, similar to pexiganan, suggests its selective membrane-disrupting activity depends on environment-specific structural transitions. These structure–activity relationships provide valuable insights for future AMP design.

Our model exhibits poor prediction performance for species with limited data. To address this issue, we can apply up-sampling and sequence perturbation, or alternatively, use transfer learning—where the model is first pretrained on data-rich species with similar characteristics and then fine-tuned using the limited data from underrepresented species. These approaches can help improve performance for species with poor data.

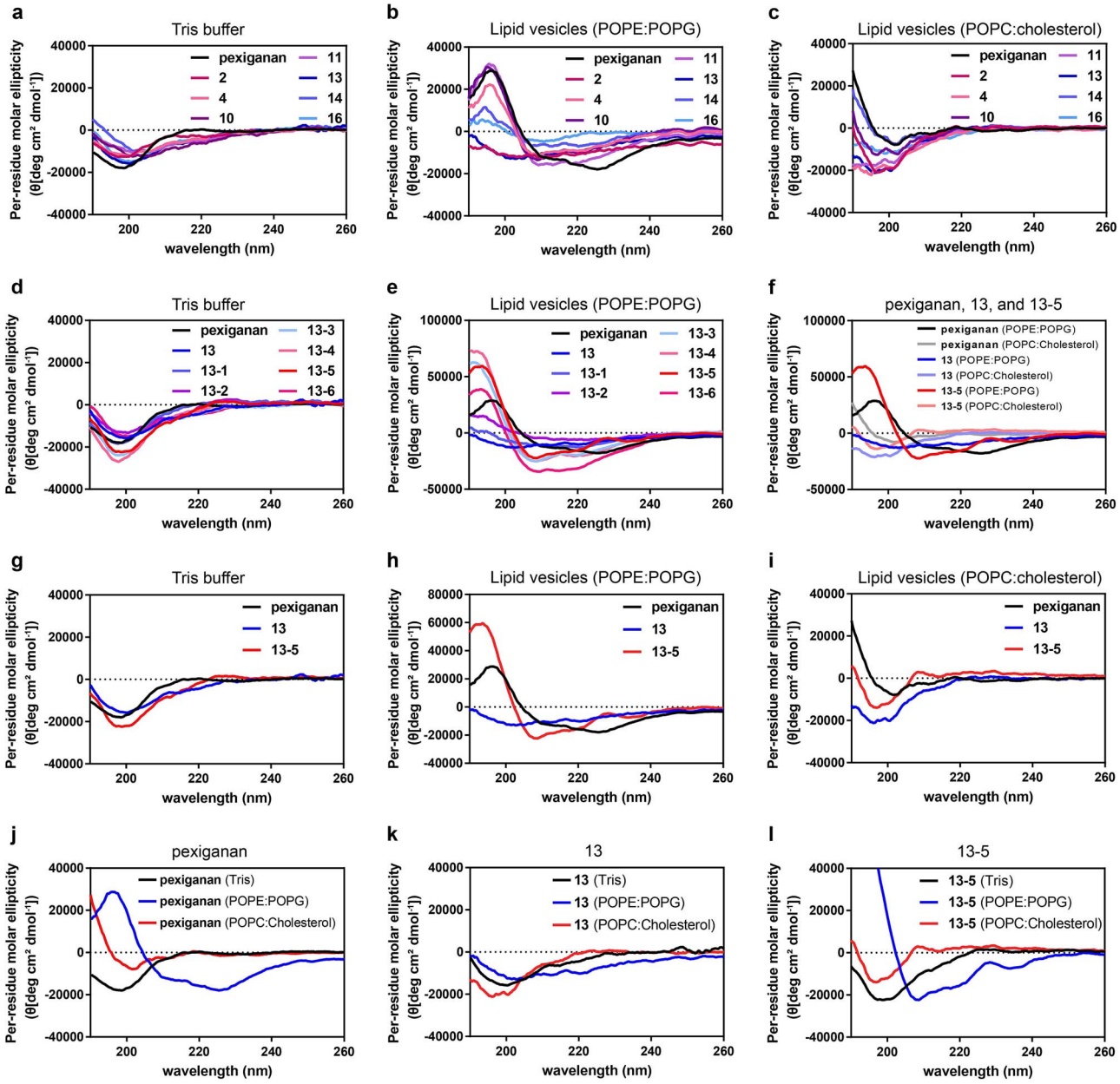

Figure 5. CD spectra of AMPs (50 μM) at 20 °C: A), d), and g) in 5 mM Tris–HCl buffer, b), e), and h) in 5 mM lipid vesicles (POPE:POPG = 7:3 in molar ratio, bacterial membrane mimic) in 10 mM Tris–HCl buffer, c) and i) in 5 mM lipid vesicles (POPC:Cholesterol = 1:1 in molar ratio, erythrocyte membrane mimic) in 10 mM Tris–HCl buffer. f) Comparison of pexiganan, 13, and 13–5 in *E. Coli* and erythrocyte-mimic lipid vesicles. Conformational changes of j) pexiganan, k) 13, and l) 13–5 in the three environments. Alt text: CD spectra comparing secondary structure of antimicrobial peptides (pexiganan, 13, and 13–5) in buffer, bacterial membrane-mimic vesicles (POPE:POPG), and erythrocyte-mimic vesicles (POPC/cholesterol), with conformational analysis across environments.

Currently, LLAMP only supports the 20 standard amino acids, limiting its ability to process non-canonical amino acids such as D-amino acids, N-methyl amino acids, peptoids, β-peptides, and post-translational modifications (PTMs). The lack of sufficient activity data for these modifications presents a significant challenge. Potential solutions include pretraining with datasets containing D-amino acids and PTMs, introducing modifiable tokens, or utilizing graph neural networks to represent modifications. These approaches could enhance the model's ability to discover improved drug-like AMPs, similar to clinically used peptides like daptomycin and polymyxin.

This study used ESM-2 as the core model. Recently released ESM-3 [38] and ESM Cambrian [39] offer different capabilities: ESM-3 specializes in controllable protein generation and could enable higher-quality AMP generation by integrating sequence, structure, and function information. ESM Cambrian improves contact precision but shows limited model size options and no significant improvement in peptide prediction performance (PPPL) as shown in Supplementary Table 8. Therefore, model selection should be carefully considered based on specific AMP prediction requirements.

While our results are promising, several areas warrant further investigation. The variation in prediction accuracy across different bacterial species suggests room for improvement in species representation during training. Additionally, the development of resistance and *in vivo* efficacy of the identified peptides

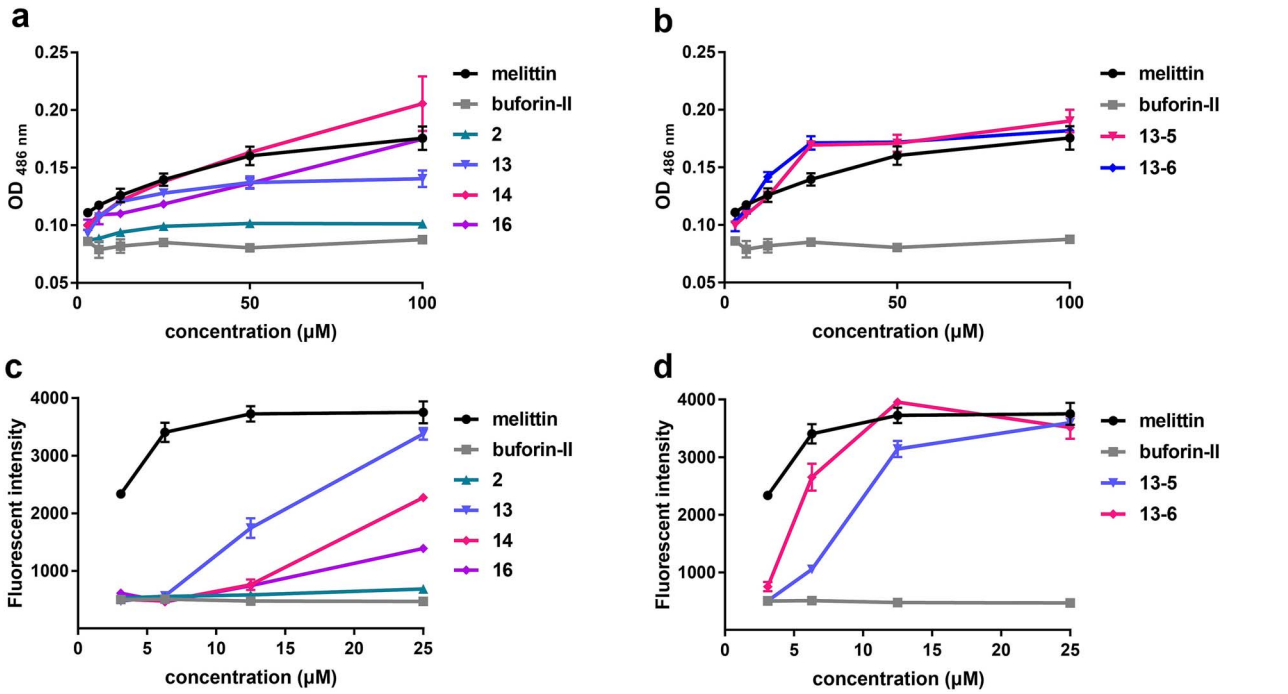

Figure 6. a) and b) effect of AMPs on bacterial outer membrane tested with the nitrocefin assays in *E. Coli* ATCC 25922. c) and d) effect of AMPs on bacterial inner membrane using PI uptake assays in *E. Coli* ATCC 25922. Alt text: Assays demonstrating antimicrobial peptide effects on *E. Coli* membranes: Nitrocefin tests for outer membrane disruption (a-b) and propidium iodide uptake tests for inner membrane disruption (c-d).

needs to be evaluated in future studies. Looking ahead, LLAMP's framework could be extended to predict activity against other pathogens, including fungi and viruses, and to optimize other peptide properties such as stability and cell penetration. The integration of additional structural and physicochemical features could further enhance prediction accuracy and guide peptide optimization.

This work establishes a powerful platform for accelerating the discovery and optimization of antimicrobial peptides, offering a promising approach to address the growing challenge of antibiotic resistance. The combination of AI-driven prediction, rational design, and experimental validation provides a robust framework for developing next-generation antimicrobial therapeutics.

Materials and methods

Datasets

The datasets utilized in this study are detailed in Table 1. Our model consists of two parts: peptide MLM tuning and regression fine-tuning. First, in the peptide MLM tuning part, the context of the peptide is learned by inputting the sequence information of the peptide. The dataset for this phase was sourced from the PeptideAtlas database (12/2022), comprising 5.5 million peptide sequences from various organisms [33]. To eliminate redundant sequences, we employed the CD-HIT program [40] with a 90% threshold, yielding 1,731,042 unique peptide sequences. Of these, 96% were allocated for training and the remaining 4% for validation.

Next, in the regression fine-tuning for the species-aware activity prediction model, we used DBAASP v3 [41] and reference genomes of bacteria obtained from NCBI [42]. The DBAASP v3 database, a manually curated resource, offers activity data for various bacterial species, including MIC values for 16 968 monomer peptide sequences targeting specific bacteria. We curated the database based on multiple criteria.

1) We selected sequences with lengths between 5 and 50 amino acids.
2) Sequences containing atypical amino acids (Z, B, J, O, U, or X) were excluded.
3) For consistency, we converted MIC values from μg/ml to μM, using the median value in cases of duplicate entries.
4) Only target species represented by at least 60 peptide sequences were considered.
5) We applied a log10 transformation to the MIC values.

Consequently, the refined dataset for fine-tuning comprises 9992 peptide sequences across 45 species. We allocated 35 of these species to the training, validation, and test sets, distributing them in an 8:1:1 ratio based on the number of peptides. Moreover, 10 species were reserved for an external set. Species-specific statistics are shown in Supplementary Tables 9 and 10.

Furthermore, we incorporated and utilized the ESKAPEE-MICpred dataset [14] and the GRAMPA (*E. coli*) dataset [13] to ensure a fair comparison of performance. Details of the comparison dataset statistics are shown in Supplementary Table 11.

Genomic features

To train the species-specific activity prediction model, we extracted features of bacterial species from the whole genome sequence downloaded from NCBI Genome (downloaded December 2022) [42]. Specifically, we computed the normalized frequencies of mono-, di-, tri-, and tetra-nucleotide compositions, yielding 4, 16, 64, and 256 features respectively, for a total of 340 genomic features per species. These k-mer frequencies were calculated using Biopython [43, 44] and normalized to account for genome size variations across different bacterial species. The resulting 340-dimensional genomic feature vector for each bacterial species

Table 7. Antimicrobial activities against pathogenic strains and hemolytic activities of peptide 13 derivatives.

Compounds	MIC[a] (μM)											HD$_{10}$/HD$_{50}$[b] (μM)	H$_{max}$[c] (200 μM)	Selectivity index[d]
	Escherichia coli ATCC 25922	*Acinetobacter baumannii* ATCC 19606	*Klebsiella pneumoniae* ATCC 700603	*Pseudomonas aeruginosa* ATCC 27853	*Salmonella typhimurium* SL 1344	*Enterobacter spp.* KCTC 2625	*Staphylococcus aureus* ATCC 25923	*Bacillus subtilis* KCTC 3068	*Enterococcus faecalis* KCTC 3511	*Listeria monocytogenes* KCTC 13064	*Enterococcus faecium* KCTC 13225	*Staphylococcus epidermidis* KCTC 1917		
melittin[e]	12.5	<3.1	12.5	12.5	6.3	<3.1	≤1.6	<3.1	3.1	nd[f]	nd[f]	nd[f]	<3.1/4.0	105.7 ± 17.6	<0.2
buforin-II[e]	12.5	50	>100	100	>100	50	12.5	50	50	nd[f]	nd[f]	nd[f]	nd[f]	nd[f]	nd[f]
omiganan[e]	12.5	12.5	25	12.5	25	<3.1	12.5	6.3	25	nd[f]	nd[f]	nd[f]	>100/>100	0	>8.0
pexiganan	6.3	<3.1	<3.1	25	6.3	<3.1	25	6.3	6.3	3.1	6.3	3.1	>100/>100	13.5 ± 1.8	>15.9
13	25	<3.1	25	>100	3.1	<3.1	50	3.1	100	50	25	25	>100/>100	0.6 ± 0.4	>4.0
13–1	50	6.3	100	100	100	50	100	50	12.5	50	>100	100	86.3/>200	33.5 ± 3.2	1.7
13–2	50	12.5	50	100	100	50	100	50	12.5	50	50	50	66.9/>200	29.9 ± 11.6	1.3
13–3	50	nd[f]	nd[f]	nd[f]	nd[f]	nd[f]	100	nd[f]	nd[f]	100	>100	100	>200/>200	0.2 ± 0.1	>4.0
13–4	50	nd[f]	nd[f]	nd[f]	nd[f]	nd[f]	100	nd[f]	nd[f]	100	25	100	>200/>200	0.3 ± 0.1	>4.0
13–5	12.5	<3.1	25	50	25	<3.1	25	50	6.3	12.5	25	6.3	>200/>200	9.3 ± 0.3	>16.0
13–6	25	6.3	50	100	50	25	25	25	12.5	12.5	25	12.5	0.6/111.7	94.8 ± 2.2	<0.1

[a]These concentrations represent mean values in triplicate. [b]HC$_{10}$ and HC$_{50}$ are the concentrations of compounds causing 10% and 50% hemolysis in rat erythrocytes, respectively. These concentrations represent mean values in triplicate. [c]Hmax is the percentage (%) of hemolysis at the highest concentration tested (200 μM). [d]The selectivity index was calculated as HC$_{10}$ divided by the minimum inhibitory concentration (MIC) values in *E. coli* ATCC 25922. [e]Literature values of melittin, buforin-II, and omiganan were obtained from ref [52]. [f]Not determine

is processed through a two-layer multilayer perceptron (MLP) to generate a compact genomic representation (Fig. 1b).

Protein language model and peptide MLM tuning

ESM-2, inspired by BERT [29], is the SOTA model among protein language models, which showed the best performance in contact prediction, and structure prediction of proteins [23]. It offers a range of model sizes, varying from 8 million to 15 billion parameters, and is readily accessible on HuggingFace [45], facilitating easy fine-tuning. For our study, we utilized the 'esm2_t12_35M-UR50D' as a baseline model, which features 12 transformer encoder layers and 35 million parameters.

To enhance the protein language model's proficiency in interpreting peptide language, we further pre-trained ESM-2 using shorter peptide sequences compared to full-length proteins (Fig. 1a). Our training process involved randomly masking 15% of the sequences. Within these masked sequences, 80% of the tokens were masked, 10% were replaced with random tokens, and the remaining 10% were left unmasked. This training technique is in line with the pre-training approaches typically used for protein language models.

Regression fine-tuning

The structure of LLAMP is shown in Fig. 1b. The LLAMP model takes a pair consisting of a peptide sequence and a bacterial genome as input. The peptide sequence passes through peptide-tuned ESM-2, generating an embedding that represents the sequence. Let $X_p \in \mathbb{N}^L$ denote the tokenized sequence of length L. We obtain contextual embeddings using the peptide-tuned ESM-2 $f_{p^{ESM}}$:

$$E = f_{p^{ESM}}(X_p) \in \mathbb{R}^{L\times 480}$$

To derive a fixed-length representation of the peptide, we select the CLS token representation $H = E_0$. The CLS token embedding $H \in \mathbb{R}^{480}$ is then linearly projected to a lower-dimensional peptide representation:

$$P = W_p H + b_p \in \mathbb{R}^{256}$$

Concurrently, the bacterial genome is transformed into a 340-dimensional vector using a genomic featurizer. Genome-derived feature vector is $G \in \mathbb{R}^{340}$. These features are passed through a two-layer MLP to produce a compact representation:

$$G_1 = ReLU(W_{g1} + b_{g1}) \in \mathbb{R}^{256}$$

$$G_2 = W_{g2} G_1 + b_{g2} \in \mathbb{R}^{128}$$

Then, the peptide and genome representations are concatenated to form a join feature vector:

$$F = \left[P \,||\, G_2\right] \in \mathbb{R}^{384}$$

These joint feature vectors are passed through a three-layer MLP:

$$F_1 = ReLU(W_1 F + b_1) \in \mathbb{R}^{256}$$

$$F_2 = ReLU(W_2 F_1 + b_2) \in \mathbb{R}^{256}$$

$$y^{pred} = W_3 F_2 + b_3 \in \mathbb{R}$$

The final prediction y^{pred} is a scalar output representing the regression target. For this, the Adam optimizer [46] and MSE loss

were used. The model is trained using MSE loss the predicted value y^{pred} and the true target value y^{label}. Given a batch of N samples, the loss is defined as:

$$\mathcal{L}_{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(y_i^{pred} - y_i^{label}\right)^2$$

Details of hyper-parameters are shown in Supplementary Table 12 and Supplementary Section 10.

PeptideAtlas screening

We aimed to discover novel AMP candidates by screening the approximately 5.5 million peptides in the PeptideAtlas database [33]. In practice, synthesizing, purifying, and evaluating all predicted peptide candidates through *in vitro* assays is not feasible due to resource limitations. Therefore, we implemented a rigorous filtering step. The database screening process was conducted as follows.

First, using our trained LLAMP model, we selected peptide sequences with predicted MIC values below 10 μM for more than 90% of either gram-positive or gram-negative bacterial species among 45 species. This step ensures that the selected candidates have the potential to be effective against a wide range of bacterial pathogens. Second, we filtered out sequences longer than 20 amino acids, a property associated with the average length of potent AMPs (Supplementary Fig. 1b). This step also eliminates longer sequences that may have limited therapeutic potential due to their size and potential challenges in synthesis and delivery. Next, to further refine our candidate list, we selected peptides with an instability index less than or equal to 40 and a net charge greater than or equal to 4 using Biopython [43]. These physicochemical properties are commonly associated with stable and effective AMPs. Finally, we refined the selection further based on helicity, hydrophobicity, and amphipathicity. The peptides were filtered by comparing their properties with those of three well-known AMPs: pexiganan, indolicidin, and LL-37. Helicity, which induces amphipathic property through appropriate alignment of hydrophobic monomers and cationic monomers within peptides, was predicted by AlphaFold2 colab [36]. Hydrophobicity was evaluated using two criteria: a hydrophobic residue composition greater than 0.4 and a GRAVY score greater than −0.990. Amphipathicity was assessed through analysis of a helical wheel diagram. Based on these criteria, seven peptides were identified.

By applying these stringent selection criteria, we were able to narrow down the vast PeptideAtlas database to a focused set of promising novel AMP candidates. The selected AMP candidates were then subjected to further *in vitro* validation to confirm their antimicrobial activity and assess their therapeutic potential.

Attention analysis and AMP sequence optimization

To gain insight into the importance of specific amino acid residues, we analyzed sequences from the fine-tuning test set. We performed attention analysis on the top 100 sequences with the highest average predictive values out of 919 test sequences. The average predicted log10 MIC values for these top 100 sequences ranged from −1.14 to 0.49.

We then input these 100 sequences into the LLAMP model to obtain the attention weights of the CLS token across all heads for each of the three fine-tuned layers. For each layer, we identified residues with attention values exceeding 0.5 and applied a hypergeometric test to each amino acid residue. This analysis allowed us to determine which amino acids were statistically overrepresented among the high-attention residues, providing insights into the model's focus during prediction.

Peptide synthesis, purification, characterization, and activity assay

The crude peptides were purchased from AnyGen Co. Ltd. (Gwangju, South Korea) or DANDI CURE Co. Ltd. (Cheongju, South Korea). Peptides were purified by preparative HPLC, and their identity and purity were confirmed using UPLC and LC–MS. UPLC chromatograms recorded at 220 nM are shown in Supplementary Figs. 8 and **9**. Calculated and observed peptide masses, determined by LC–MS, are provided in Supplementary Table 13. Detailed procedures for peptide synthesis, purification, characterization, and activity assays are described in Section 11 and Section 12 of the **Supplementary Information**.

Key Points

- Large Language model for antimicrobial peptide (AMP) activity prediction represents an innovative approach to AMP discovery, integrating species-specific information with protein language modeling to accurately predict antimicrobial potency.
- Through computational screening of over 5.5 million peptide sequences followed by in vitro validation, our model successfully identified promising candidates with distinctive properties: peptide **13** with the highest selectivity, peptide **16** with highest potency.
- By analyzing the attention mechanisms within our model, we identified specific amino acids (e.g., Trp, Lys, and Phe) that significantly contribute to antimicrobial activity, providing rational design principles for peptide optimization.

Author contributions

D.B., M.K., J.S., H. N. (Conceptualization), D.B., M.K. (Data curation), (Formal analysis), (Investigation), (Methodology), (Software), (Validation), (Visualization), (Writing - original draft), and J.S., H.N. (Funding acquisition), (Investigation), (Methodology), (Project administration), (Resources), (Supervision), (Writing - review & editing)

Supplementary data

Supplementary data are available at *Briefings in Bioinformatics* online.

Additional file 1. Supplementaryinformation

The supplementary description on Introduction, Materials, Methods, and additional Results with figures and tables.

Additional file 2. Supplementary Data 1 and 2

Acknowledgements

We appreciate the high-performance GPU computing support of HPC-AI Open Infrastructure via GIST SCENT.

Conflict of interest: None declared.

Funding

This work was supported by the Gwangju Institute of Science and Technology (GIST) research fund (Future-leading Specialized Research Project, 2025), and supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2024-00344200, RS-2024-00401422, RS-2023-00217317), and supported by a grant of the Korea Machine Learning Ledger Orchestration for Drug Discovery Project (K-MELLODDY), funded by the Ministry of Health & Welfare and Ministry of Science and ICT, Republic of Korea (grant number: RS-2024-00460010).

Data availability

The processed datasets used in this study are available on the GitHub repository (https://github.com/GIST-CSBL/LLAMP).

The raw datasets are all publicly available.

DBAASP version 3: https://dbaasp.org/

PeptideAtlas: https://peptideatlas.org/

GRAMPA: https://github.com/zswitten/Antimicrobial-Peptides/blob/master/data/grampa.csv

ESKAPEE-MICpred: https://eskapee-micpred.anvil.app/

Code availability

The LLAMP source code, trained model parameters, and scripts are available on the GitHub repository (https://github.com/GIST-CSBL/LLAMP).

References

1. Reardon S. WHO warns against 'post-antibiotic' era. *Nature* 2014. https://doi.org/10.1038/nature.2014.15135.
2. Uddin TM, Chakraborty AJ, Khusro A. *et al.* Antibiotic resistance in microbes: History, mechanisms, therapeutic strategies and future prospects. *J Infect Public Health* 2021;**14**:1750–66. https://doi.org/10.1016/j.jiph.2021.10.020.
3. O'neill, J. Antimicrobial resistance: Tackling a crisis for the health and wealth of nations. *Rev Antimicrob Resist* 2014.
4. Prestinaci F, Pezzotti P, Pantosti A. Antimicrobial resistance: A global multifaceted phenomenon. *Pathog Glob Health* 2015;**109**:309–18. https://doi.org/10.1179/2047773215Y.0000000030.
5. Malkoski M, Dashper SG, O'Brien-Simpson NM. *et al.* Kappacin, a novel antibacterial peptide from bovine milk. *Antimicrob Agents Chemother* 2001;**45**:2309–15. https://doi.org/10.1128/AAC.45.8.2309-2315.2001.
6. Lai Y, Villaruz AE, Li M. *et al.* The human anionic antimicrobial peptide dermcidin induces proteolytic defence mechanisms in staphylococci. *Mol Microbiol* 2007;**63**:497–506. https://doi.org/10.1111/j.1365-2958.2006.05540.x.
7. Maróti Gergely G, Kereszt A, Kondorosi É. *et al.* Natural roles of antimicrobial peptides in microbes, plants and animals. *Res Microbiol* 2011;**162**:363–74. https://doi.org/10.1016/j.resmic.2011.02.005.
8. Lee I, Keum J, Nam H. DeepConv-DTI: Prediction of drug-target interactions via deep learning with convolution on protein sequences. *PLoS Comput Biol* 2019;**15**:15. https://doi.org/10.1371/journal.pcbi.1007129.
9. Lee I, Nam H. Sequence-based prediction of protein binding regions and drug–target interactions. *J Chem* 2022;**14**:5. https://doi.org/10.1186/s13321-022-00584-w.
10. Bae B, Bae H, Nam H. LOGICS: Learning optimal generative distribution for designing de novo chemical structures. *J Chem* 2023;**15**:77. https://doi.org/10.1186/s13321-023-00747-3.
11. Yan J, Bhadra P, Li A. *et al.* Deep-AmPEP30: Improve short antimicrobial peptides prediction with deep learning. *Mol Ther Nucleic Acids* 2020;**20**:882–94. https://doi.org/10.1016/j.omtn.2020.05.006.
12. Li C, Sutherland D, Hammond SA. *et al.* AMPlify: Attentive deep learning model for discovery of novel antimicrobial peptides effective against WHO priority pathogens. *BMC Genomics* 2022;**23**:77. https://doi.org/10.1186/s12864-022-08310-4.
13. Witten J, Witten Z. Deep learning regression model for antimicrobial peptide design. *BioRxiv* 2019;692681.
14. Sharma R, Shrivastava S, Singh SK. *et al.* Artificial intelligence-based model for predicting the minimum inhibitory concentration of antibacterial peptides against ESKAPEE pathogens. *IEEE J Biomed Health Inform* 2024;**28**:1949–58. https://doi.org/10.1109/JBHI.2023.3271611.
15. Veltri D, Kamath U, Shehu A. Deep learning improves antimicrobial peptide recognition. *Bioinformatics* 2018;**34**:2740–7. https://doi.org/10.1093/bioinformatics/bty179.
16. Bárcenas O, Pintado-Grima C, Sidorczuk K. *et al.* The dynamic landscape of peptide activity prediction. *Comput Struct Biotechnol J* 2022;**20**:6526–33. https://doi.org/10.1016/j.csbj.2022.11.043.
17. Krizhevsky A, Sutskever I, Hinton GE. ImageNet classification with deep convolutional neural networks. *Commun ACM* 2017;**60**:84–90. https://doi.org/10.1145/3065386.
18. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997;**9**:1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.
19. Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Netw* 2005;**18**:602–10. https://doi.org/10.1016/j.neunet.2005.06.042.
20. Vaswani A, Shazeer N, Parmar N. *et al.* Attention is all you need. *Adv Neural Inf Process Syst* 2017;**30**:5999–6009.
21. Elnaggar A, Heinzinger M, Dallago C. *et al.* ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Trans Pattern Anal Mach Intell* 2022;**44**:7112–27. https://doi.org/10.1109/TPAMI.2021.3095381.
22. Rives A, Meier J, Sercu T. *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci* 2021;**118**:e2016239118. https://doi.org/10.1073/pnas.2016239118.
23. Lin Z, Akin H, Rao R. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 1979;**379**:1123–30. https://doi.org/10.1126/science.ade2574.
24. Zhang Y, Lin J, Zhao L. *et al.* A novel antibacterial peptide recognition algorithm based on BERT. *Brief Bioinform* 2021;**22**:bbab200. https://doi.org/10.1093/bib/bbab200.
25. Dee W. LMPred: Predicting antimicrobial peptides using pre-trained language models and deep learning. *Bioinformatics Advances* 2022;**2**:vbac021. https://doi.org/10.1093/bioadv/vbac021.
26. Lee H, Lee S, Lee I. *et al.* AMP-BERT: Prediction of antimicrobial peptide function based on a BERT model. *Protein Sci* 2023;**32**:e4529. https://doi.org/10.1002/pro.4529.

27. Salazar J, Liang D, Nguyen TQ. *et al.* Masked language model scoring. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* 2020:2699–2712. https://doi.org/10.18653/v1/2020.acl-main.240.
28. McInnes L, Healy J, Melville J. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv*:1802.03426, 2018.
29. Devlin J, Chang M-W, Lee K. *et al.* Bert: pre-training of deep bidirectional transformers for language understanding. *In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies*; **1**, 4171–86.
30. Lawrence TJ, Carper DL, Spangler MK. *et al.* amPEPpy 1.0: A portable and accurate antimicrobial peptide prediction tool. *Bioinformatics* 2021;**37**:2058–60. https://doi.org/10.1093/bioinformatics/btaa917.
31. Cao Q, Ge C, Wang X, *et al.* Designing antimicrobial peptides using deep learning and molecular dynamic simulations. *Brief Bioinform* 2023;**24**:bbad058. https://doi.org/10.1093/bib/bbad058.
32. Wang R, Wang T, Zhuo L. *et al.* Diff-AMP: Tailored designed antimicrobial peptide framework with all-in-one generation, identification, prediction and optimization. *Brief Bioinform* 2024;**25**:bbae078. https://doi.org/10.1093/bib/bbae078.
33. Deutsch EW, Lam H, Aebersold R. PeptideAtlas: A resource for target selection for emerging targeted proteomics workflows. *EMBO Rep* 2008;**9**:429–34. https://doi.org/10.1038/embor.2008.56.
34. Guruprasad K, Reddy BVB, Pandit MW. Correlation between stability of a protein and its dipeptide composition: A novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Engineering, Design and Selection* 1990;**4**:155–61. https://doi.org/10.1093/protein/4.2.155.
35. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 1982;**157**:105–32. https://doi.org/10.1016/0022-2836(82)90515-0.
36. Mirdita M, Schütze K, Moriwaki Y. *et al.* ColabFold: Making protein folding accessible to all. *Nat Methods* 2022;**19**:679–82. https://doi.org/10.1038/s41592-022-01488-1.
37. Monteiro C, Pinheiro M, Fernandes M. *et al.* A 17-mer membrane-active MSI-78 derivative with improved selectivity toward bacterial cells. *Mol Pharm* 2015;**12**:2904–11. https://doi.org/10.1021/acs.molpharmaceut.5b00113.
38. Hayes T, Rao R, Akin H. *et al.* Simulating 500 million years of evolution with a language model. *Science* 2025;**387**:850–8. https://doi.org/10.1126/science.ads0018.
39. ESM Team. ESM Cambrian: Revealing the mysteries of proteins with unsupervised learning. *EvolutionaryScale Website*. 2024; https://evolutionaryscale.ai/blog/esm-cambrian.
40. Fu L, Niu B, Zhu Z. *et al.* CD-HIT: Accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**:3150–2. https://doi.org/10.1093/bioinformatics/bts565.
41. Pirtskhalava M, Amstrong AA, Grigolava M. *et al.* DBAASP v3: Database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic Acids Res* 2021;**49**:D288–97. https://doi.org/10.1093/nar/gkaa991.
42. O'Leary NA, Wright MW, Brister JR. *et al.* Reference sequence (RefSeq) database at NCBI: Current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* 2016;**44**:D733–45. https://doi.org/10.1093/nar/gkv1189.
43. Cock PJA, Antao T, Chang JT. *et al.* Biopython: Freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* 2009;**25**:1422–3. https://doi.org/10.1093/bioinformatics/btp163.
44. Takahashi M, Kryukov K, Saitou N. Estimation of bacterial species phylogeny through oligonucleotide frequency distances. *Genomics* 2009;**93**:525–33. https://doi.org/10.1016/j.ygeno.2009.01.009.
45. Wolf T, Debut L, Sanh V. *et al.* HuggingFace's transformers: state-of-the-art natural language processing. *arXiv preprint*. 2020; https://doi.org/10.48550/arXiv.1910.03771.
46. Kingma DP, Ba J. Adam: A method for stochastic optimization. *arXiv preprint*. 2017; https://doi.org/10.48550/arXiv.1412.6980.
47. Gautier R, Douguet D, Antonny B. *et al.* HELIQUEST: A web server to screen sequences with specific α-helical properties. *Bioinformatics* 2008;**24**:2101–2. https://doi.org/10.1093/bioinformatics/btn392.
48. Breiman L. Random forests. *Machine Learning* 2001;**45**:5–32.
49. Dubchak I, Muchnik I, Holbrook SR. *et al.* Prediction of protein folding class using global description of amino acid sequence. *Proc Natl Acad Sci* 1995;**92**:8700–4. https://doi.org/10.1073/pnas.92.19.8700.
50. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 2016, 785–94.
51. Cortes C, Vapnik V. Support-vector networks. *Machine Learning* 1995;**20**:273–97.
52. Kim M, Cheon Y, Shin D. *et al.* Real-time monitoring of multitarget antimicrobial mechanisms of Peptoids using label-free imaging with optical diffraction tomography. *Advanced Science* 2023;**10**:2302483. https://doi.org/10.1002/advs.202302483.

Briefings in Bioinformatics, 2025, **26(4)**, bbaf343
https://doi.org/10.1093/bib/bbaf343
Problem Solving Protocol