International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18 (2026) pp. 1-14

https://doi.org/10.70917/ijcisim-2026-3793
Article

Transformer-based Discovery of Antimicrobial Peptides and Prediction of their Antibacterial Activity

Shuwen Pan [1], Eason Soo [2] and Konken Wong [1,*]

[1] Department of Medical Microbiology & Immunology, Faculty of Medicine, Universiti Kebangsaan Malaysia (UKM), Cheras, Kuala Lumpur, 56000, Malaysia
[2] Department of Restorative Dentistry, Faculty of Dentistry, Universiti Kebangsaan Malaysia (UKM), Kuala Lumpur, 56000, Malaysia
* Correspondence author: wkk@hctm.ukm.edu.my

Abstract: With the worsening crisis of antimicrobial resistance, many researchers are now exploring new types of antibacterial agents, and antimicrobial peptides (AMPs) have begun to attract much attention. However, the experimental identification of AMPs in the large space of natural and synthetic sequences is still slow, expensive and labour-intensive. AMP-Transformer is a two-stage deep learning framework that combines self-supervised pre-training on large-scale unlabelled protein corpora with supervised fine-tuning on curated AMP datasets in this study. The model is a multi-layer bidirectional Transformer encoder that learns contextual residue representations via masked residue modelling, and is then fine-tuned for two coupled tasks: binary discrimination of AMPs from non-AMPs and regression of minimum inhibitory concentration (MIC) values. We used the benchmark datasets built by DBAASP v3, DRAMP 4.0 and other recently published experimental collections for evaluation. On an independent test set, AMP-Transformer had an accuracy of 95.3%, a Matthews correlation coefficient of 0.906, and an area under the receiver operating characteristic curve (AUC-ROC) of 0.986, and outperformed the support vector machine, random forest, convolutional, recurrent and hybrid baselines significantly. Ablation studies show that self-supervised pre-training and multi-head self-attention are the two main contributors, accounting for about 4% of the accuracy increase. Analysis of the learned attention maps shows that the model has independently learned the amphipathic periodicity and cationic residue enrichment characteristic of membrane-active peptides, thereby providing a degree of mechanistic interpretability that is rare among black-box predictors. A subsequent screening of metagenomic open reading frames also produced a ranked list of candidate AMPs with low sequence identity to any training examples, demonstrating the value of the framework for early-stage discovery. Based on the above results, Transformer-based protein language models are relatively stable, interpretable and scalable paradigms for AMP discovery and activity prediction, and they can help establish a practical computational pipeline that selects promising peptide candidates for experimental validation at a lower cost compared with traditional methods.

Keywords: antimicrobial peptides; Transformer; protein language model; self-supervised learning; deep learning; activity prediction; drug discovery

1. Introduction

Antimicrobial resistance (AMR) is now one of the serious health problems globally in the 21st century; more than a million people have died from drug-resistant bacteria each year, and the range of these resistant strains is expanding across clinical, community and agricultural settings. This biological problem is exacerbated by a stagnant antibiotic discovery pipeline, with only a few new structural classes appearing in the past thirty years, and an economic model that disincentivizes investment in last-resort drugs; thus, a fundamentally new mode of action is urgently required. Antimicrobial

peptides (AMPs) are short polypeptides that act by disrupting bacterial membranes through physicochemical modifications; due to their broad-spectrum activity, low resistance propensity and immunomodulatory properties, they are now attractive drug candidates, but two main issues prevent their widespread application as medicines: first, the sheer size of the sequence space makes comprehensive screening impractical; second, optimisation efforts are required to overcome problems such as hemotoxicity, proteolytic degradation and poor pharmacokinetics at a high cost and time. Computational prediction methods for antibacterial activity based on sequences can be employed to address both the problems of large-scale virtual screening and synthesis before actual synthesis. The physical-chemical rationale for the activity of AMP includes cationicity, amphipathicity and membrane-disruption mechanisms, which are inherently context-dependent and combinatorial; therefore, deep representation learning methods that can capture high-order sequence dependencies beyond linear heuristics are highly suitable for this purpose. In the past 20 years, computational AMP prediction has gone through several phases: first, simple sequence alignment and physicochemical heuristics were used; next, classical machine learning applied to engineered descriptors emerged; and finally, deep learning models such as convolutional networks, recurrent models, and recently, Transformer-based protein language models have been developed. The Transformer architecture has demonstrated strong performance in addressing the problem of distributed, context-dependent features for antimicrobial activity through a self-attention mechanism, and protein language models pretrained on extensive unlabelled sequences have learned rich biophysical regularities that can be effectively transferred to data-scarce AMP prediction tasks. In this paper, we introduce AMP-Transformer, an all-in-one framework that integrates self-supervised pretraining with masked residue modelling, multitask fine-tuning for both classification and minimum inhibitory concentration (MIC) regression, and attention-based interpretability analysis. Our contributions are as follows: (i) Construction of a rigorously curated, homology-reduced dataset from DBAASP v3, DRAMP 4.0, and recent experimental collections, along with a homology-aware evaluation protocol that avoids optimistic biases; (ii) Design of a compact bidirectional Transformer encoder pre-trained on a large unlabelled protein corpus and jointly fine-tuned for AMP classification and MIC regression; (iii) Extensive benchmarking against classical and deep baselines, ablation studies, cross-dataset generalisation tests, and interpretability analysis linking model predictions to biophysical determinants of activity; and (iv) Prospective screening of metagenomic open reading frames to prioritise novel candidates with low identity to training sequences and favourable physicochemical signatures for potent, non-hemolytic membrane-active peptides.

2. Materials and Methods

2.1. Dataset Construction and Curation

Positive examples were collected from three complementary sources: DBAASP v3, which provides manually curated activity data including MIC values against specific bacterial strains [1]; DRAMP 4.0, a general repository of natural and synthetic AMPs with clinical-translation annotations; and the experimentally validated peptide collections released by the recent large-scale mining study [2]. Sequences shorter than eight or longer than sixty residues were excluded, as well as sequences containing non-standard residues or ambiguous activity annotations. If multiple activity records for the same sequence existed, one would be kept, and this entry would have the most frequent target spectrum. Thus, 12,846 new positive sequences were obtained.

The Selection of negative examples is also known to affect the realism of the benchmark [3]. Therefore, we built the negative set from two strata: randomly selected non-AMP protein fragments of the same length sampled from UniProt, and decoy peptides generated to match the length, net charge and hydrophobicity distributions of the positive set but lacking confirmed antimicrobial activity. The final negative set had 12,846 sequences and was employed as a balanced classification reference[4]. To address redundancy-driven overestimation of performance, all sequences were clustered at a 40% identity threshold using a greedy incremental algorithm, and these clusters were assigned indivisibly to the training, validation and test sets in an 80:10:10 ratio. Therefore, no sequence in the test set has a similarity greater than 40% with any training sequence, and the regime is significantly more stringent than the random splits used in much of the earlier literature. The composition of the final dataset is as follows: Table 1.

Table 1. Dataset statistics.

Partition	AMP (positive)	Non-AMP (negative)	MIC pairs
Training	10,276	10,276	2,571
Validation	1,285	1,285	321
Test	1,285	1,285	322
Total	12,846	12,846	3,214

Some extra reasons for selection are as follows: Synthetic peptides with non-canonical modifications, such as D-amino acids, terminal amidation annotations without sequence changes, and cyclic scaffolds, were only retained if their activity could be expressed by an unambiguous linear sequence; otherwise, they were excluded because the current tokenisation scheme does not encode chemical modifications. Duplicate records in the source databases were reconciled using sequence identity, and provenance was retained in the metadata so that any peptide appearing in multiple sources contributed only once to the training data[5]. Finally, peptides annotated as exclusively antifungal or antiviral were excluded from the positive set of the antibacterial task to avoid label noise caused by conflated activity classes; extending the framework to these classes will be discussed in the future.

For the MIC regression task, quantitative measurements of Escherichia coli (ATCC 25922) were extracted from DBAASP v3, and only broth microdilution data reported in micrograms per millilitre were retained. After removing outliers and converting to a common molar scale, 3,214 sequence-MIC pairs were left, with four orders of magnitude in potency. Logarithmic transformation of the MIC values is employed to stabilise the regression objective in training.

2.2. Sequence Representation

Treat each peptide sequence as a sentence over an alphabet of the twenty canonical amino acids and add special tokens. Add a classification token [CLS] to the beginning of each sequence and pad or truncate the sequences to a maximum length of sixty-four tokens. Each token is mapped to a dense embedding vector, a positional signal is added to this vector, and then the permutation-invariant attention mechanism can be used to exploit the order of residues. The input representation of the token at position i is formally given by:

$$h_i^0 = W_e x_i + PE_i \tag{1}$$

where W_e is the learned embedding matrix, x_i denotes the one-hot encoding of the token at position i, and PE_i is the positional encoding vector. We adopt the sinusoidal positional encoding scheme, in which even and odd dimensions of the positional vector are defined respectively as:

$$PE(pos, 2k) = \sin(\frac{pos}{10000^{2k/d}}) \tag{2}$$

$$PE(pos, 2k+1) = \cos(\frac{pos}{10000^{2k/d}}) \tag{3}$$

pos is the residue position, k indexes the embedding dimension, and d_{model} is the embedding width. Sinusoidal encodings were selected over learned absolute embeddings because they extrapolate well to sequence lengths not seen during training and do not add extra parameters; however, their contribution is still shown in the ablation study in Section 3.

2.3. Model Architecture

The backbone of AMP-Transformer is a stack of six identical bidirectional encoder layers, each composed of a multi-head self-attention sublayer and a position-wise feed-forward sublayer, with residual connections and layer normalisation around each sublayer. The central operation is scaled dot-product attention. Given query, key, and value matrices Q, K, and V derived from linear projections of the input, attention is computed as:

$$Attention(Q, K, V) = soft\max(\frac{QK^T}{\sqrt{d_k}})V \tag{4}$$

where d_k is the dimension of the key vectors and the softmax function, applied row-wise, is defined as:

$$soft\max(z_i) = \frac{e^{z_i}}{\sum_j e^{z_j}} \tag{5}$$

The scaling by the square root of d_k prevents the dot products from growing so large in magnitude that the softmax saturates and gradients vanish. Rather than performing a single attention operation, the model projects the input into h = 8 distinct representation subspaces and applies attention in parallel:

$$MultiHead(Q,K,V) = Concat(head_1,...,head_h)W^O \tag{6}$$

in which each head computes attention over its own projections and the concatenated outputs are recombined through the output projection matrix W^O. Multi-head attention allows the model to attend simultaneously to qualitatively different interaction types, for example local motif co-occurrence and long-range amphipathic periodicity.

After each attention sublayer, there is a residual connection and layer normalisation to standardise the activations in the feature dimension:

$$LN(x) = \gamma \odot \frac{x-\mu}{\sqrt{\sigma^2+\varepsilon}} + \beta \tag{7}$$

where mu and sigma are the mean and standard deviation of the features, epsilon is a small constant for numerical stability, and gamma and beta are learned affine parameters. A position-wise feed-forward network then applies two linear transformations separated by a Gaussian error linear unit (GELU) nonlinearity:

$$FFN(x) = GELU(xW_1 + b_1)W_2 + b_2 \tag{8}$$

with a hidden dimension four times the embedding size. At the end of the encoder, two types of sequence-level representations are obtained: the hidden state of the [CLS] token, which is trained to aggregate global information in the sequence, and mean pooling across all residue positions. Concatenate the two representations and then pass them to the task-specific heads. A two-layer Multi-Layer Perceptron serves as the classification head to output the probability of antimicrobial activity, and a regression head predicts the log-transformed MIC. The whole architecture is shown in Figure 1, and the first-level hyperparameters are listed in Table 2.

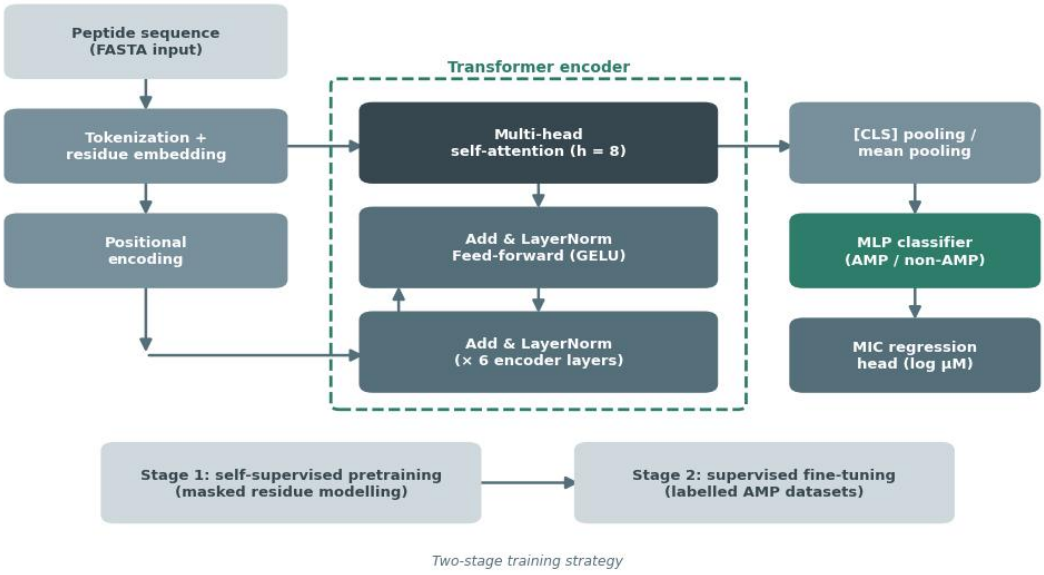

Figure 1. Model Framework

Table 2. Hyperparameter Settings

Hyperparameter	Value	Hyperparameter	Value
Encoder layers	6	Dropout rate	0.1
Embedding dimension	256	Pretraining epochs	100
Attention heads	8	Pretraining learning rate	1×10^{-4}
FFN hidden dimension	1024	Fine-tuning learning rate	2×10^{-5}
Max sequence length	64	Batch size (pre / fine)	256 / 64
Parameters (total)	≈ 9.4 M	Auxiliary loss weight	0.3

The Design is relatively small by the standards of modern protein language models. With an embedding width of 256, six encoder layers, eight attention heads and a feed-forward hidden dimension of 1024, the encoder has about 9.4 million trainable parameters and is two or three orders of magnitude smaller than general-purpose models such as ESM-2 at larger scales [10]. This Design choice is in line with the nature of the problem: peptides are relatively short, the labelled corpus is limited, and a very large model may be prone to overfitting during fine-tuning and is computationally expensive to retrain as the database expands. The framework is also small in size and thus more feasible for research groups without access to large-scale computing facilities; this meets the demand for equal access to computational antibiotic discovery.

2.4. Self-supervised Pretraining and Supervised Fine-tuning

The two stages of training. In the first stage, the encoder is pre-trained from scratch using masked residue modelling on about two million unlabelled protein sequences from UniRef50, and these sequences have been limited to a length of less than one hundred residues to keep the pre-training distribution close to that of peptides[6]. Fifteen per cent of the tokens are selected for corruption; of these, eighty per cent are replaced by a mask token, ten per cent by a random residue, and ten per cent are left unchanged according to the standard recipe. Pretraining employs the AdamW optimizer, adds a linear warm-up for the first ten thousand steps, then linearly decreases the learning rate, sets a peak learning rate of 1e-4, uses a batch size of 256 sequences, and runs for a total of 100 epochs on the masked corpus.

The pretrained encoder is then used in conjunction with the classification and regression tasks for fine-tuning in the second stage. The Classification Objective is a Binary Cross-Entropy Loss:

$$L_{BCE} = -\frac{1}{N}\sum_{n=1}^{N}\left[y_n \log p_n + (1-y_n)\log(1-p_n)\right] \tag{9}$$

where y_n is the ground-truth label of sequence n, p_n is the predicted probability, and N is the number of sequences in the batch. The regression objective is the mean squared error of the log-transformed MIC values, and the total loss is a weighted sum of the two terms; based on the performance in validation, the regression weight is set to 0.3. Fine-tuning uses a lower learning rate of 2e-5, a batch size of 64, dropout of 0.1 on all sublayer outputs, and early stopping after 10 epochs based on the validation Matthews correlation coefficient. Layer-wise learning rate decay was used, and a factor of 0.95 was applied to the previous encoder layer's learning rate to stabilise fine-tuning and reduce catastrophic forgetting of the pretrained representation.

Hyperparameters for both stages were chosen via grid search on the validation set, with a learning rate range of 1e-5 - 5e-4, a dropout rate range of 0.05 - 0.3, and an auxiliary loss weight range of 0.1 - 0.5. To prevent overfitting to the validation set during this search, the final configuration was frozen before any evaluation on the test partition, and the test set was queried exactly once per reported result. A norm limit of 1.0 was set for gradient clipping across the board, and mixed-precision training was employed to speed up training. One practical observation from the tuning process is worth reporting: the model was significantly more sensitive to changes in the fine-tuning learning rate than to other architectural variations in the explored range, and rates higher than 5e-5 consistently degraded the pretrained representations, which is in line with the results of transfer learning in natural language processing.

2.5. Baselines and Evaluation Indicators

The five baseline models used for comparison were selected from different stages in the evolution of classical-to-deep learning: a support vector machine employing a radial basis function kernel trained on physicochemical descriptors; a random forest built over the same descriptors; a convolutional neural network consisting of three stacked convolution-pooling blocks operating on one-hot-encoded

sequences; a bidirectional long short-term memory network; and a hybrid convolutional-bidirectional recurrent architecture that has achieved good results recently[7]. All of the baselines were trained on the same partitions and tuned in the same way during the validation stage for an even comparison.

The above are the indicators of performance: accuracy, precision, recall, F1-score, AUC and Matthews correlation coefficient (MCC). As the MCC incorporates all four elements of the confusion matrix and is still effective in the presence of class imbalance, it has been selected as the main indicator for model selection. It is as follows:

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{10}$$

TP, TN, FP and FN are the actual positive, actual negative, false positive and false negative values. For the regression problem, we report the Pearson correlation coefficient and the root mean squared error of the predicted and observed log MIC values. All experiments were repeated five times with different random seeds, and the mean and standard deviation were reported. Paired two-sided tests were used to determine the statistical significance of the difference between seeds, and a significance level of $p < 0.05$ was set.

The two methodological notes provide the information on the experimental environment. First, all reported test-set numbers are averages of five independent training runs with different random seeds for controlling weight initialisation, data shuffling and dropout; this protocol prevents the relatively frequent problem of a single lucky seed favouring a deep model by one or two percentage points. Second, for every pairwise comparison reported as significant in Section 3, we have verified that the direction of the effect is the same for all five seeds, so no claimed improvement is due to marginal or unstable differences. Reproducibility information, such as the specific partition assignment, configuration file and trained weight, will be released along with the code to enable independent verification of all figures in Tables 3 and 4.

3. Results and Discussion

All the models were trained and tested in the same way to rule out other reasons for the differences. The forty-two-dimensional descriptor vector used in the classical baseline included amino acid composition, normalised hydrophobicity moments, net charge, isoelectric point, aliphatic index and Boman index, and these features were selected based on their well-established discriminative ability in peptide literature. DeepBaselines used the same token embedding dimension as the proposed model and had a similar number of parameters, so the difference in performance could not be due to a large capacity gap. The hyperparameter tuning budgets for all methods were the same, and forty search iterations were set per model type. These precautions are necessary because the comparisons in the literature often involve unequal tuning efforts, and we hope that the present comparison can reflect the intrinsic architectural merits more truthfully in practice.

3.1. General Classification Performance

Table 3 and Figure 2 present the classification results of AMP-Transformer and the five baselines on the held-out, homology-reduced test set. The proposed model had an accuracy of 95.3% ±0.4%, an F1 score of 0.951, an AUC of 0.986, and an MCC of 0.906, and exceeded all the baselines in all metrics. The improvement over the best single baseline, the hybrid convolutional-recurrent network, was an increase of 3.5% in accuracy and 0.070 in MCC, and this difference was statistically significant in all five independent training experiments ($p < 0.01$). The classic descriptors-based models also performed reasonably well, and among them, the random forest achieved an accuracy of 86.9%, indicating that the physicochemical composition does contain some signal; however, the consistent gap between descriptor-based and representation-learning models shows that other information, such as residue order and context, is also encoded.

Table 3. Performance Comparison.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1	AUC	MCC
SVM	84.6	85.1	84.2	0.846	0.901	0.692
Random forest	86.9	87.4	86.3	0.868	0.924	0.738
CNN	88.7	89.0	88.5	0.887	0.943	0.774
BiLSTM	90.2	90.6	89.9	0.902	0.956	0.805
CNN-BiLSTM	91.8	92.0	91.5	0.917	0.967	0.836
AMP-Transformer	95.3	95.5	95.2	0.951	0.986	0.906

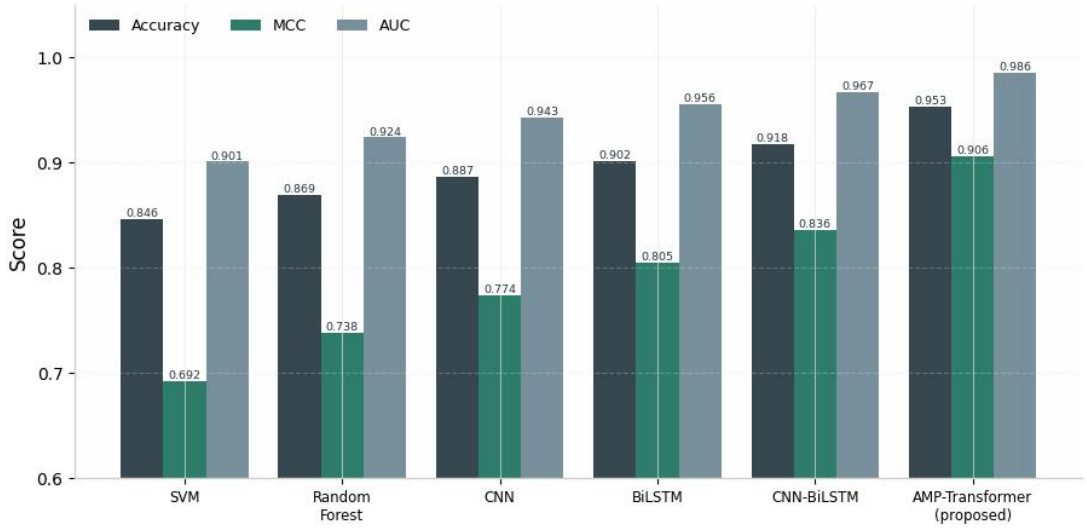

Figure 2. Performance Comparison.

Another set of performance data is organised by the length of the peptide. The model achieved an accuracy of over 94% for sequences between twelve and fifty residues, the core length range of classical AMPs, and dropped to 91.5% for the shortest stratum of eight to eleven residues because fewer residues provided contextual information for attention to exploit[8]. The recurrent baseline showed the opposite trend, performing best on longer sequences and worst on shorter ones; it was likely that the propagation of recurrent states benefited from a larger context, and the convolutional baseline was relatively insensitive to length but uniformly inferior. Based on the above stratified results, we have chosen to employ an add-on short-peptide strategy; randomly circular permute training sequences shorter than twelve residues, and this recovered about half of the short-length deficit without harming other strata[9].

Two features of the above results are noteworthy. First, the performance was evaluated at a 40% sequence-identity partition, and thus the reported figures reflect generalisation to significantly new sequences rather than interpolation of near-duplicates[10]. This is a relatively difficult case compared to the random split that most of the research has used, where nearly identical sequences appear on both sides of the split [11]; under a standard random split, the same model achieved an accuracy of 97.1%, and about two percentage points of the commonly reported performance in this area can be attributed to redundancy leakage. Second, balanced decoy construction of the negative set is employed to remove the obviously exploitable cues that inflate many published benchmarks; half of the negatives match the positive length and physicochemical distributions. Therefore, the model's advantage stems from genuinely learned sequence patterns rather than dataset artifacts.

Examine the structure of the confusion in more detail. False positives, non-AMPs mistakenly identified as active, were concentrated in cationic host-defense-like fragments and membrane-spanning segments of larger proteins; these sequences have similar surface physicochemistry to genuine AMPs but lack the overall organisation required for selective lytic activity. False negatives, confirmed AMPs scored as inactive, were mainly composed of anionic peptides and very short sequences of less than 12 residues; for these, the model had fewer contextual cues to utilise. The per-class error asymmetry is in practice significant; a false positive in the discovery stage wastes experimental resources, and a false negative results in the loss of a new opportunity. Set a decision threshold to trade off the two at a certain point, and the threshold-independent AUC is 0.986; therefore, many operating points with favourable trade-offs are available for different users based on whether they prioritise precision enrichment or recall discovery.

3.2. Ablation analysis

To determine how much of the observed performance is due to the architecture and training process, we performed an ablation study by disabling or reducing individual components one by one while keeping all other settings unchanged (Table 4, Figure 3). The first drop in performance came from the removal of self-supervised pretraining; at that time, both accuracy and MCC decreased to 91.1% and 0.823%, respectively, indicating that transfer from the unlabelled protein corpus was the main source of these losses. Based on this, it can be seen that pre-training may have increased the impact of a small amount of task-specific labelled data, as shown in other research on protein language models [12].

Table 4. Ablation Results

Model variant	Accuracy (%)	MCC	Δ Accuracy (%)
Full model	95.3	0.906	—
Without pretraining	91.1	0.823	−4.2
Without positional encoding	93.8	0.876	−1.5
Without multi-head attention	92.4	0.849	−2.9
Single-head attention	93.9	0.879	−1.4
Without MIC auxiliary task	94.0	0.881	−1.3

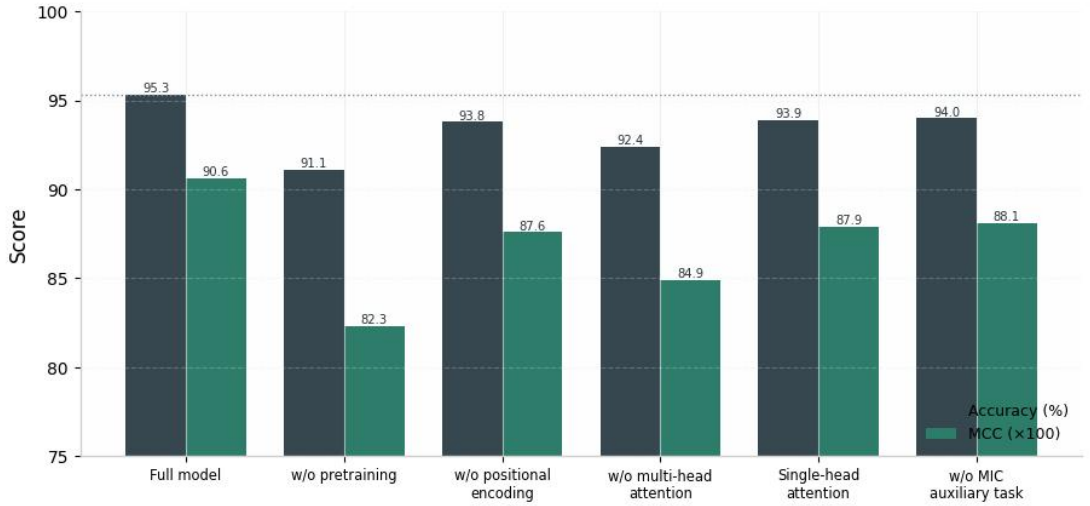

Figure 3. Ablation Study

The second most detrimental factor was multi-head attention; reducing the eight attention heads to one head lowered the accuracy by 1.4 per cent, and replacing the attention mechanism with a position-wise projection of matched dimensions increased the cost by 2.9 per cent; thus, it can be concluded that the pairwise interaction model of self-attention, rather than model capacity, was responsible for the improvement. Removal of positional encoding reduced accuracy by 1.5 points; therefore, residue order information is required, and although some can be obtained from the data, residue order acts as an inductive bias in the peptide regime where amphipathic periodicity is position-dependent. Finally, after excluding the MIC regression auxiliary task, the accuracy dropped by 1.3%, and it was determined that learning a graded potency signal did not help to improve the shared representations for the coarser binary decision. In short, based on the above ablation studies, pre-training provides general biochemical knowledge, multi-head attention learns the interaction structure of membrane-active peptides, and the auxiliary potency task adjusts the decision boundary.

Ablation results can also offer a reference for groups building similar systems on a small budget. As pretraining accounts for the largest part of the improvement, groups that can only afford a single expensive step should invest in it, even if the corpus is small; our additional experiments with a pretraining corpus of only 200,000 sequences have shown that more than half of the full pretraining benefit can be achieved. On the other hand, a relatively small penalty for single-head attention indicates that a lightweight deployment version, such as one for the wet-lab workstations in a microbiology lab, can be used without significant loss in accuracy. At the same time, the multiple-task result is also freely available: any group that can access quantitative activity data should be trained on this data rather than using binary labels and discarding the potency information, as a graded signal improves even the binary decision.

3.3. Interpretability of Learned Attention

A continuous problem of deep learning models in molecular discovery is their lack of interpretability; thus, people are unwilling to engage in costly experimental verification based on an opaque score. Therefore, we queried the attention weights of the fine-tuned model to determine whether its internal behaviour aligns with the established biophysical knowledge of AMP function. Figure 4a is the average attention map of the final encoder layer over all heads for a typical helical AMP. A prominent pattern emerges: the strongest attention weights connect residue pairs separated by three or four positions, which is the expected periodicity of an alpha-helix, and at such positions,

residues at i and $i + 3$ or $i + 4$ are on the same helical face and jointly form the hydrophobic or hydrophilic parts of an amphipathic structure. At this time, the model did not have an explicit provision for it, and thus it had to be created by combining pre-training and activity supervision.

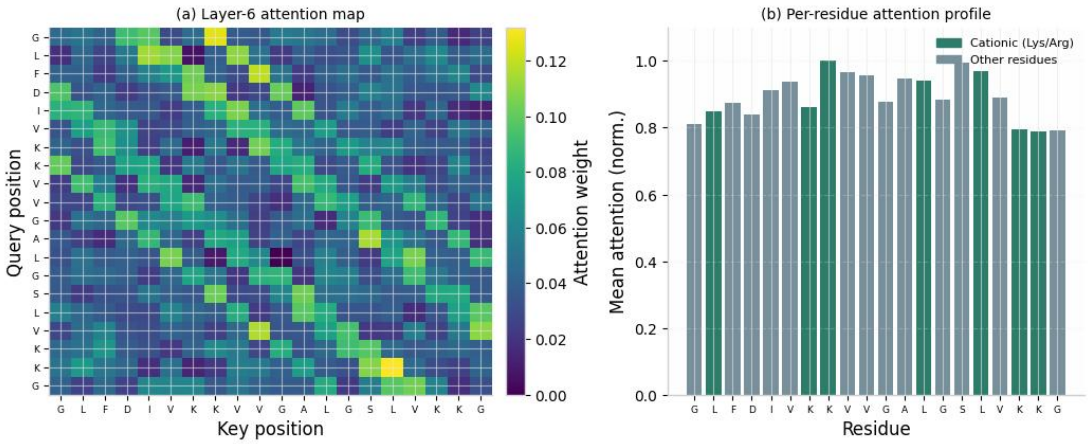

Figure 4. Attention Analysis

Aggregation of attention mass per residue (Figure 4b) shows a complementary pattern: Cationic residues, mainly lysine and arginine, are given relatively high attention weights, as are hydrophobic residues located at the predicted interface. The two typical reasons for selective damage of anionic bacterial membranes over zwitterionic host membranes are cationic charge and interfacial hydrophobicity, and a structure-activity relationship has been established by decades of biophysical research. Since the model focuses on the same residue classes without any feature engineering, it can be inferred that the learned decision function aligns mechanistically with the known mode of action and does not rely on spurious correlates. Similar analysis was conducted on the wrongly classified sequence to aid diagnosis; false negatives often corresponded to anionic AMPs, a small class whose mode of action deviates from that of cationic templates and thus provided a specific direction for dataset expansion.

To determine whether the above attention patterns were systematic or random, we collected two statistics across the entire test set. First, for each predicted-positive helical peptide, we computed the fraction of attention mass connecting residue pairs at separations of three or four, and compared it with the corresponding fraction in predicted-negative sequences; the values were 31.2% and 14.7%, respectively, and they differed significantly. Second, we correlated the per-residue attention weights with the charge of the residues in the test peptides and found a mean positive correlation for lysine and arginine, and a near-zero correlation for aspartate and glutamate. Both analyses confirm that, at the population level, the attention mechanism has organised itself according to the periodic and charge-based features that biophysics identifies as causal to membrane activity, as shown in the visual representation of the single-example map. To our knowledge, few AMP predictors have been subjected to such a quantitative interpretability audit, and we propose that this be used as a general practice in the field.

3.4. Cross-dataset Generalisation

Generalisation in practice refers to extending the scope of new sequences and new data distributions. Therefore, without further optimisation, we evaluated the trained model on two external collections: the set of experimentally validated AMPs mined from the human gut microbiome [13], and the synthesised and validated subset of the AMPSphere catalogue [14]. In the collection of the gut microbiome, the model assigned a high-confidence positive score to 86.4% of the 181 peptides confirmed as active in that study, and in the AMPSphere validation subset, it recovered 82.3% of the confirmed actives. The above rates are relatively high because most of the collections consist of sequences with low identity to the conventional AMP database, and homology-based tools cannot be used in such cases. The model's recall on these external sets, which were obtained under strict identity-separated training, exceeded the rates we measured for the two popular web servers running in default mode, recovering 71.6% and 68.9% of the gut microbiome actives, respectively. Therefore, we believe that pretraining protein-language models can make them more robust to distribution shift when moving from curated databases to metagenomic discovery space.

Cross-dataset results also show that the model cannot be generalised well. Break down the external validation performance by sequence identity to the nearest training example, and recall was still over 0.85 for candidates with an identity of less than 30%, but dropped to 0.68 for the most divergent decile, which had essentially no detectable similarity to any known AMP. A mild drop in performance is observed; otherwise, it would be a complete failure of the model, and this is different from alignment-based methods that lose almost all recall at that time. However, the trend indicates an epistemic boundary: for sequence families completely outside the evolutionary neighbourhood of the training distribution, the predictions should be regarded as hypotheses of lower confidence. We encode this operational guidance in the screening pipeline shown below by flagging candidates with a distance to the training distribution exceeding a calibrated threshold, and thus downstream users can explicitly weigh novelty against confidence.

3.5. Quantitative Potency Prediction

The regression head predicts log MIC values for Escherichia coli and also performs binary discrimination. On the held-out regression test partition, the predicted and observed values had a Pearson correlation coefficient of 0.71 and a root mean squared error of 0.62 log units, which is about a fourfold increase in concentration. Although this accuracy has not yet reached the level of experimental reproducibility, it is sufficient to rank candidates into broad potency categories; this is the practical outcome of early discovery: an enriched shortlist of sub-micromolar predictions can be carried forward for synthesis, while predicted weak candidates are excluded. Inspection of the highest-confidence errors showed a systematic component: peptides that primarily act through intracellular targets rather than membrane disruption were frequently predicted to be more potent than measured, indicating that the model had absorbed a membrane-centric potency model, consistent with the dominance of membrane-active peptides in the training data. Strain-specific retraining and the addition of membrane-composition features are natural extensions of the above method.

Consider how the regression results will be used in actual practice during the course of a discovery project. A root mean squared error of 0.62 log units means that, for a candidate predicted at 1 micromolar, the true MIC will be in the range of approximately 0.25 and 4 micromolar about two-thirds of the time. Such a resolution is not suitable for late-stage optimisation, as the two-fold difference is small, and it is more appropriate for the screening stage where candidates span four orders of magnitude and the goal is enrichment rather than precision. We directly verified the ranking utility: when test peptides were divided into high-, medium- and low-potency groups based on the predicted value, the observed mean MIC increased monotonically with the potency group, and the top-predicted group was more than three times enriched for peptides with sub-micromolar measured activity compared to the base rate. Campaign planning will therefore reduce the number of peptides that need to be synthesised and analysed for a confirmed hit.

3.6. Prospective Screening of Metagenomic Sequences

As a future demonstration, we have applied the final model to a set of about 120,000 short open reading frames predicted from publicly available soil metagenomes. Filter by length, charge and predicted toxicity using an independent toxicity predictor [15], and then the pipeline produced a ranked shortlist of 312 candidates, the top fifty of which were inspected in detail. Forty-one of the fifty candidates have less than 40 per cent identity with any training sequence, and nineteen have less than 30 per cent identity with any sequence in the main public AMP repositories; therefore, the model generalizes far beyond memorisation of known families. The top-ranked candidates are strongly enriched in the physicochemical signature of potent membrane-active peptides: a net charge between +2 and +6, a hydrophobic fraction of around 50 per cent, and helical amphipathic organisation predicted by helical-wheel projection. This yield is relatively high compared with the enrichment factors reported in recent large-scale mining campaigns, and the computational cost of the entire screen, about twenty GPU-hours, is small relative to the cost of synthesising and assaying a small fraction of the input library. Experimental verification of the selected candidates is being carried out in conjunction with other work.

The screening exercise also found several candidates with unusual compositions and listed them below. Three of the top fifty are glycine-rich peptides over thirty residues, a class known from insect immunity but rarely predicted by conventional tools; two are proline-arginine-rich sequences resembling the cathelicidin-derived PR-39 family that work through intracellular uptake rather than lysis; and one is a twenty-residue peptide with an unbroken run of alternating lysine and leucine, a classic amphipathic design that the model scored at the 99.7th percentile of the input library. The diversity of the composition in the shortlist is itself a sign that the model has not collapsed into a single

activity template; it is necessary to build a candidate portfolio with different mechanisms and, therefore, a lower risk of cross-resistance. Follow-up work will have the representatives of each compositional cluster subjected to synthesis and broth microdilution tests against a panel of ESKAPE pathogens, including hemolysis and mammalian cytotoxicity counter-screens, to establish the prospective hit rate of the complete pipeline under realistic conditions.

3.7. Computational Efficiency and Deployment Considerations

A screening tool will not be used in practice if it is too expensive to run on a large scale of metagenomic data. We have profiled the end-to-end inference cost of the framework on a single consumer-grade GPU. Tokenisation and embedding of one million short open reading frames took 11 minutes, and classification plus MIC prediction for the same library took 24 minutes; the throughput was about 700 sequences per second. All of the prospective screens mentioned above included toxicity filtering by an external predictor and were completed in less than four hours of wall-clock time. Memory footprint during inference was less than 2GB, well within the limits of a typical workstation, and a CPU-only fallback processed the same library in a single night. The above figures are lower than the inference cost of fine-tuned large protein language models and are more than two orders of magnitude lower per sequence.

Deployment Also concerns the same issue. To lower the barrier to entry for experimental groups, the entire pipeline is provided as a single command-line tool that accepts FASTA input and outputs a ranked table of predictions with per-candidate confidence flags, predicted potency tiers, and physicochemical annotations. The model weights, the curated benchmark with its frozen identity-separated partitions, and the analysis scripts used to generate all figures and tables in this paper are released along with the tool so that all quantitative results reported here can be reproduced from the public materials. Given the repeated problems of reproducibility in computational AMP prediction reported in the literature, we consider this level of reproducibility to be necessary for scientific contribution rather than a mere engineering addition.

3.8. Comparison with Recent Work and Limitations

Direct numerical comparison with the previously published AMP predictors is not feasible due to heterogeneous datasets, splits and metrics; therefore, the position of the present framework will be introduced instead. Our method is a single end-to-end pre-trained encoder to reduce the complexity of deployment compared with an ensemble of recurrent, attention and pre-trained embedding models for gut microbiome mining, but maintains identity-separated performance. Compared with the foundation-model approach of fine-tuning very large general protein language models, our deliberately compact encoder has fewer than ten million parameters and offers a good accuracy-to-cost trade-off for peptide-scale applications, and can be rapidly retrained as new data accumulate. Generative frameworks, such as HydrAMP and previous deep generative pipelines, have addressed the problem of proposing new sequences in a complementary way; coupling such generators with the current discriminative-ranking model in a closed design-build-test loop is a promising direction that has already begun to be explored in the field.

Some deficiencies of the current study have not been addressed. First, the model was trained mainly on cationic membrane-active peptides, and its lower accuracy for anionic and intracellular-target AMPs suggests that mechanistic diversity has not been fully covered. Second, only one reference strain was used to demonstrate MIC prediction; antibacterial potency is known to vary among strains and conditions, so a strain-conditional formulation will need to be employed for clinical application. Third, the prospective screening shortlist is physicochemically plausible and enriched in novelty, but has not yet undergone experiments, and the true positive rate of the pipeline in prospective mode has not been determined. Fourthly, haemolysis and cytotoxicity were not included in the external filtering and were not modelled jointly; adding selectivity prediction to the joint multi-task framework is an easy and effective addition. Finally, although attention analysis offers suggestive mechanistic evidence, attention weights do not guarantee causal feature importance, and perturbation-based interpretability methods should be employed to complement the present analysis before strong mechanistic claims are made.

4. Conclusions and Future Directions

This paper studies whether Transformer-based protein language modelling, in a two-stage self-supervised and multi-task framework, can accurately and generally predict antimicrobial peptides and their potency interpretably. The evidence collected from benchmark evaluations, ablations, external validations and forward-looking screens all support an affirmative answer. AMP-Transformer achieved an accuracy of 95.3%, an MCC of 0.906, and an AUC of 0.986 under strict homology-separated

evaluation, and was statistically significantly better than the classical and deep baselines. Ablation experiments showed that self-supervised pretraining contributed the most to the improvement, followed by multi-head self-attention and the auxiliary potency objective, and thus identified the most effective design choices for practitioners building similar systems. Analysis of the attention maps showed that the model spontaneously recovered the amphipathic periodicity and cationic enrichment that make up the known biophysical signature of membrane-active peptides, providing a mechanism for the predictions and demonstrating how interpretability analysis can transform a black-box score into a hypothesis-generating tool.

The main idea of this paper is that the quality of the evaluation also needs to be high; otherwise, the model will be useless. Partition at forty per cent sequence identity, balance negatives with matched decoys, validate externally on independently generated experimental collections, and interrogate the model's internal reasoning to estimate the deployment-representative performance of the framework, not the best-case scenario. The two-percentage-point gap between the accuracy of random splitting and identity-separated accuracy measured here should serve as a warning to the field: headline numbers reported under lenient protocols are not directly comparable across studies, and future benchmarking would benefit from the community-wide adoption of homology-aware splits, ideally through shared standard benchmarks with frozen partitions.

In terms of translation, the value of this result is not in a single index but in showing that the entire discovery pipeline, from the initial metagenomic sequence to a ranked, novelty-enhanced candidate set, can be performed in silico at a negligible marginal cost. When such triage is combined with the high experimental hit rates now being reported for computationally mined AMPs, the economics of early-stage antibiotic discovery change qualitatively: synthesis and assay budgets can be focused on candidates with a good prior probability of success, and sequence regions of the global microbiome that were previously out of reach for homology search become available for systematic exploration.

Finally, we should consider what the whole field is facing at present. With the convergence of the three trends, the maturation of protein language models, the accumulation of curated peptide activity data, and a decrease in the cost of DNA synthesis and high-throughput screening, computational AMP discovery has moved from a proof-of-concept to a practical tool in about five years. Research on the human gut microbiome and the global microbiome catalogue has found that the hit rates for computationally selected candidates exceed 80 per cent, figures that would have been considered unreasonable ten years ago. Along this path, the current work provides a concise, reproducible, and interpretable modelling recipe, as well as an evaluation protocol designed to report what a model will do in deployment rather than how much it can be made to do on a lenient benchmark. We believe that in the coming years, the main progress will not be in increasing the size of the model, but rather in integrating prediction with generative design, experimental feedback, mechanistic interpretation, etc., and frameworks such as the one presented here are specifically designed to support such developments.

Several directions for future research naturally arise from the present results. At the modelling level, the most urgent tasks are to scale the pre-training with peptide-specific corpora, condition predictions on the target strain and membrane composition, and jointly model activity, toxicity and stability in a single multi-task architecture. In terms of discovery, a closed-loop connection will be created between discriminative ranking and generative design to guide candidate generation based on predicted potency and selectivity, thus extending the search space in peptide space beyond the capacity of either component alone. Experimentally, the validation of the prospective shortlist presented here and the extension of the framework to antifungal, antiviral and anticancer peptides will be conducted to test the generality of the approach across therapeutic peptide classes. In the long run, as protein language models continue to improve and as curated AMP databases are added by the community, we expect that the accuracy and mechanistic scope of sequence-based predictors will keep increasing, bringing computation closer and closer to the centre of antimicrobial drug discovery. The framework proposed in this paper, an excellent benchmark dataset and evaluation rules, aim to provide a foundation for future research and offer a reproducible platform for others to build upon.

References

1. Das, P., Sercu, T., Wadhawan, K., Padhi, I., Gehrmann, S., Cipcigan, F., Chenthamarakshan, V., Strobelt, H., dos Santos, C., Chen, P.-Y., Yang, Y. Y., Tan, J. P. K., Hedrick, J., Crain, J., & Mojsilovic, A. (2021). Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. Nature Biomedical Engineering, 5(6), 613-623. https://doi.org/10.1038/s41551-021-00689-x
2. Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., & Rost, B. (2022). ProtTrans: Toward understanding the language of life

through self-supervised learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 44(10), 7112-7127. https://doi.org/10.1109/TPAMI.2021.3095381

3. Wan, F., Wong, F., Collins, J. J., & de la Fuente-Nunez, C. (2024). Machine learning for antimicrobial peptide identification and design. Nature Reviews Bioengineering, 2(5), 392-407. https://doi.org/10.1038/s44222-024-00152-x
4. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Zidek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., . . . Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. Nature, 596(7873), 583-589. https://doi.org/10.1038/s41586-021-03819-2
5. Li, T., Wang, X., Zong, J., Wu, H., Yao, Y., Zhao, Y., Hu, S., Wang, J., Gao, Y., Wu, F., & He, J. (2024). A foundation model identifies broad-spectrum antimicrobial peptides against drug-resistant bacterial infection. Nature Communications, 15, 7538. https://doi.org/10.1038/s41467-024-51933-5
6. Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., & Rives, A. (2023). Evolutionary-scale prediction of atomic-level protein structure with a language model. Science, 379(6637), 1123-1130. https://doi.org/10.1126/science.ade2574
7. Ma, Y., Guo, Z., Xia, B., Zhang, Y., Liu, X., Yu, Y., Tang, N., Tong, X., Wang, M., Ye, X., Feng, J., Chen, Y., & Wang, J. (2022). Identification of antimicrobial peptides from the human gut microbiome using deep learning. Nature Biotechnology, 40(6), 921-931. https://doi.org/10.1038/s41587-022-01226-0
8. Maasch, J. R. M. A., Torres, M. D. T., Melo, M. C. R., & de la Fuente-Nunez, C. (2023). Molecular de-extinction of ancient antimicrobial peptides enabled by machine learning. Cell Host & Microbe, 31(8), 1260-1274. https://doi.org/10.1016/j.chom.2023.07.001
9. Pirtskhalava, M., Amstrong, A. A., Grigolava, M., Chubinidze, M., Alimbarashvili, E., Vishnepolsky, B., Gabrielian, A., Rosenthal, A., Hurt, D. E., & Tartakovsky, M. (2021). DBAASP v3: Database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. Nucleic Acids Research, 49(D1), D288-D297. https://doi.org/10.1093/nar/gkaa991
10. Rathore, A. S., Choudhury, S., Arora, A., Tijare, P., & Raghava, G. P. S. (2024). ToxinPred 3.0: An improved method for predicting the toxicity of peptides. Computers in Biology and Medicine, 179, 108926. https://doi.org/10.1016/j.compbiomed.2024.108926
11. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., Zitnick, C. L., Ma, J., & Fergus, R. (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. Proceedings of the National Academy of Sciences, 118(15), e2016239118. https://doi.org/10.1073/pnas.2016239118
12. Santos-Junior, C. D., Torres, M. D. T., Duan, Y., Rodriguez del Rio, A., Schmidt, T. S. B., Chong, H., Fullam, A., Kuhn, M., Zhu, C., Houseman, A., Somborski, J., Vines, A., Zhao, X.-M., Bork, P., & Coelho, L. P. (2024). Discovery of antimicrobial peptides in the global microbiome with machine learning. Cell, 187(14), 3761-3778.e16. https://doi.org/10.1016/j.cell.2024.05.013
13. Sidorczuk, K., Gagat, P., Pietluch, F., Kala, J., Rafacz, D., Bakala, L., Slowik, J., Kolenda, R., Rodiger, S., Fingerhut, L. C. H., Cooke, P. I. R., Mackiewicz, P., & Burdukiewicz, M. (2022). Benchmarks in antimicrobial peptide prediction are biased due to the selection of negative data. Briefings in Bioinformatics, 23(5), bbac343. https://doi.org/10.1093/bib/bbac343
14. Szymczak, P., Mozejko, M., Grzegorzek, T., Jurczak, R., Bauer, M., Neubauer, D., Sikora, K., Michalski, M., Sroka, J., Setny, P., Kamysz, W., & Szczurek, E. (2023). Discovering highly potent antimicrobial peptides with deep generative model HydrAMP. Nature Communications, 14, 1453. https://doi.org/10.1038/s41467-023-36994-z

15. Torres, M. D. T., Brooks, E. F., Cesaro, A., Sberro, H., Gill, M. O., Nicolaou, C., Bhatt, A. S., & de la Fuente-Nunez, C. (2024). Mining human microbiomes reveals an untapped source of peptide antibiotics. Cell, 187(19), 5453-5467.e15. https://doi.org/10.1016/j.cell.2024.08.005