nature communications

Article

https://doi.org/10.1038/s41467-024-51933-2

A Foundation Model Identifies Broad-Spectrum Antimicrobial Peptides against Drug-Resistant Bacterial Infection

Received: 28 January 2024

Accepted: 20 August 2024

Published online: 30 August 2024

Check for updates

Tingting Li[1,2,12], Xuanbai Ren[3,12], Xiaoli Luo[3], Zhuole Wang[1,2], Zhenlu Li[4], Xiaoyan Luo[3], Jun Shen[1,2], Yun Li[5], Dan Yuan[1,2], Ruth Nussinov[6,7], Xiangxiang Zeng[3] ✉, Junfeng Shi[1,2] ✉ & Feixiong Cheng[8,9,10,11] ✉

Development of potent and broad-spectrum antimicrobial peptides (AMPs) could help overcome the antimicrobial resistance crisis. We develop a peptide language-based deep generative framework (deepAMP) for identifying potent, broad-spectrum AMPs. Using deepAMP to reduce antimicrobial resistance and enhance the membrane-disrupting abilities of AMPs, we identify, synthesize, and experimentally test 18 T1-AMP (Tier 1) and 11 T2-AMP (Tier 2) candidates in a two-round design and by employing cross-optimization-validation. More than 90% of the designed AMPs show a better inhibition than penetratin in both Gram-positive (i.e., *S. aureus*) and Gram-negative bacteria (i.e., *K. pneumoniae* and *P. aeruginosa*). T2-9 shows the strongest antibacterial activity, comparable to FDA-approved antibiotics. We show that three AMPs (T1-2, T1-5 and T2-10) significantly reduce resistance to *S. aureus* compared to ciprofloxacin and are effective against skin wound infection in a female wound mouse model infected with *P. aeruginosa*. In summary, deepAMP expedites discovery of effective, broad-spectrum AMPs against drug-resistant bacteria.

Antimicrobial resistance is one of the most serious public health threats at present, leading to an estimated 10 million deaths per year by 2050[1]. Antibiotics are the most widely used medicines for multiple infectious diseases worldwide. Owing to misuse and overuse of antibiotics, drug-resistant bacteria, such as Methicillin-resistant *Staphylococcus aureus* (MRSA), have emerged and spread rapidly, resulting in significant antimicrobial resistance crisis[2]. Despite the urgency, fewer new antibiotics were developed and approved by the U.S. Food and Drug Administration (FDA), and the last entirely original one was discovered in the late 1980s[3]. In addition, a new antibiotic agent was estimated to cost around 1.5 billion dollars and take more than 10 years from basic research discovery to FDA-approved medicine[4]. It is thus urgent to develop new technologies and approaches to identify broad-spectrum, potent antibiotics, in particular for antimicrobial peptides (AMPs), in order to address antimicrobial resistance crisis and effective treatment of infectious diseases.

AMPs are typically composed of 10-50 amino acids and are considered alternative antibiotic candidates for tracking antibiotic

[1]Affiliated Hospital of Hunan University, School of Biomedical Sciences, Hunan University, Changsha, China. [2]Greater Bay Area Institute for Innovation, Hunan University, Guangzhou 511300 Guangdong Province, China. [3]College of Information Science and Engineering, Hunan University, Changsha, China. [4]School of Life Science, Tianjin University, Tianjin 300072, China. [5]Department of Ophthalmology, The 2nd Xiangya Hospital of Central South University, Changsha, China. [6]Computational Structural Biology Section, Frederick National Laboratory for Cancer Research in the Laboratory of Cancer Immunometabolism, National Cancer Institute, Frederick, MD 21702, USA. [7]Department of Human Molecular Genetics and Biochemistry, Sackler School of Medicine, Tel Aviv University, Tel Aviv, Israel. [8]Cleveland Clinic Genome Center, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA. [9]Genome Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA. [10]Department of Molecular Medicine, Cleveland Clinic Lerner College of Medicine, Case Western Reserve University, Cleveland, OH, USA. [11]Case Comprehensive Cancer Center, School of Medicine, Case Western Reserve University, Cleveland, OH, USA. [12]These authors contributed equally: Tingting Li, Xuanbai Ren. ✉e-mail: xzeng@hnu.edu.cn; Jeff-Shi@hnu.edu.cn; chengf@ccf.org

resistance because of their membrane disruption capacity[5]. Through their unique antimicrobial mechanism-of-action, AMPs have shown potency in treatment of resistant bacteria with high clinical benefits[6,7]. Penetratin has good cell permeability and average overall antimicrobial activity[8,9], and its derivatives have been used as a universal delivery vehicle[10,11]. However, traditional experimental approaches to identify AMPs are highly time-consuming and expensive[12,13]. Developing new technologies for rapid identification of effective, broad-spectrum AMPs for effective treatment of various infectious diseases to address the global antimicrobial resistance health threat is thus a pressing need.

Advanced machine learning models have been developed for the discovery of peptides[14–21]. Classic methods, such as virtual screening[15,17] and de novo designed molecules, have been applied toward discovery of potential AMPs[16,18,19]. However, there are several potential challenges for these classic methods: (1) The efficacy of virtual screening methods is profoundly influenced by the quality of the underlying data and the methods employed for feature extraction. Furthermore, current methods for screening antimicrobial peptides exhibit limitations in terms of generalization and adaptability; (2) Unconstrained generation methods are prone to disadvantages such as difficult peptide synthesis and high experimental costs; and (3) Despite the availability of experimental AMP data, its scale is modest compared to the large datasets in natural language processing. This size disparity presents challenges for traditional deep learning models in adequately learning the chemical and biological properties crucial for antimicrobial activity from these smaller datasets.

Inspired by the success of language models in natural language processing, we developed deepAMP, a peptide language model requiring limited resources for discovering potential AMPs with high potency (Fig. 1). This model takes peptides with low antimicrobial activity as inputs and output analogs with high antimicrobial activity and broad-spectrum resistance. To address the issue of scarce training resources, we propose a pre-training and multiple fine-tuning strategy, augmenting data through a sequence degradation approach. At a microscopic level, we pre-trained a generalized peptide generative model (deepAMP-general) in an unsupervised manner from a large and varied sequence database of peptides, making it capable of non-functional peptide generation. Subsequently, the dataset was expanded through a sequence degradation approach with deepAMP-general to construct AMP pairs. In a two-stage fine-tuning process, we fine-tuning the AMP optimization model (deepAMP-AOM) to identify highly potent AMP candidates, and re-fine-tuning the penetratin optimization model (deepAMP-POM) to reduce antimicrobial resistance by enhancing the membrane disruption capacity of candidate AMPs. As proof of concept, our experimental assays show that more than 90% of the candidate sequences achieve better inhibitory outcome at lower concentrations than the original AMPs. The ability to rapidly discover antimicrobial peptides indicates that the combination of original AMPs and deep language model optimization can effectively identify AMPs with broad-spectrum antimicrobial effects.

Results

Overview of deepAMP

In order to discover potential AMPs with broad-spectrum antibacterial activity and reduced risk of drug resistance. A 16-mer peptides, penetratin was chosen as the template sequence because of its good cell permeability and average overall antimicrobial activity. The optimization comprises four sub-processes (Fig. 1): (1) a deepAMP-general (pre-training model) that is responsible for rational peptide generation, (2) deepAMP-AOM (fine-tuning model) that optimize low antibacterial activity peptides and (3) deepAMP-POM (re-fine-tuning model) that optimize for penetratin, and (4) deepAMP-predict for screening prior to experimental assays.

For the first step, we formulate the problem as a mask language model by pre-training on 300,000 peptide sequences obtained from UniProt with sequence lengths between 10 and 50 (see Methods). DeepAMP-general learns the syntax rules of the peptide sequences and is able to generate non-functional peptides that are valid, diverse and not in the training set. Owing to the small number of penetratin sequences available (29 sequences), we propose a sequence degradation approach that employs deepAMP-general to transform existing highly bioactive peptides into multiple low activity peptide thereby constructing AMP pairs. This approach leverages deepAMP-general's capability to generate non-functional peptides and significantly expands our training dataset, resulting in the creation of 1009 penetratin pairs. For the second step, we fine-tuned deepAMP-AOM by AMP pairs to learn the activity characteristics key for converting low and high activity AMPs, akin to how a language model translates different languages. For the third step, we used the penetratin pairs re-fine-tuning model to make deepAMP-POM learn the characteristics of high antimicrobial activity and cell permeability. Prior to the experiment, we scored the candidate AMPs using deepAMP-predict (for more details, see Methods).

Optimization of antimicrobial peptides

To evaluate our method, we chose an existing bioactive AMP of Temporin-Ali[21] and four existing bioactive fragments of Pg-AMP1[22] as optimization templates. Temporin-Ali was optimized and experimentally validated across three rounds of iterations[21] (termed Baseline-T). The performance of deepAMP in optimizing the AMPs was compared to four methods: Random mutation, Baseline-T[21], PepCVAE[23] and HydrAMP[16]. Random mutation is a random change of an amino acid site in a random mask sequence to any other amino acid. Baseline-T is the optimized AMP using a genetic algorithm[21]. PepCVAE is one of the state-of-the-art approaches to peptide generation using the conditional variational autoencoder (CVAE) framework[23]. HydrAMP is an AMP optimization method using the CVAE[16]. Our presented deepAMP-AOM is a generic antimicrobial peptide optimization model, which is similar to the antimicrobial peptide pre-training model. We used an experimentally reported training set[21] on deepAMP-AOM for fine-tuning to obtain deepAMP-TOM, and used deepAMP-TOM for Temporin-Ali optimization task comparison. All details are provided in the Methods section and Supplemental Methods.

Here we conducted 3 rounds of iterations for optimization. The candidate AMPs generated in each iteration are evaluated using well-established scoring matrix[21] and deepAMP-predict (Fig. 2a–d). We found that the average results after optimization of all methods except deepAMP-TOM in the first iteration were lower than the scores of the original sequence (Fig. 2a). Although the average scores of all methods are higher than the original sequence after 3 iterations, the deepAMP-TOM score is optimal (Fig. 2b, c). Furthermore, the scoring results of deepAMP-predict showed that deepAMP-TOM has the best AMP optimization ability (Fig. 2d). The exceptional performance of deepAMP in optimizing AMPs demonstrated its efficiency, accuracy and ability to identify potential AMPs.

In the Pg-AMP1 fragment optimization task, Porto et al.[22] employed a genetic algorithm to optimize the four fragments of Pg-AMP1 peptide through 400 rounds of iterations and scored using a fitness function (*cf.* Methods), which achieved an optimal fitness score of approximately 0.53. The publicly available dataset includes only the top 100 guavanin candidates generated (fitness score ranging from 0.245 to 0.393) from the first 100 rounds of iterations. Therefore, we used these experimentally validated datasets as the baseline (termed Baseline-G) for our comparisons. We benchmarked the performance of deepAMP against four different methods: Random Mutation, Baseline-G, PepCVAE, and HydrAMP. For a fair comparison, we used the same training set proposed by Porto et al. on deepAMP-AOM for fine-tuning

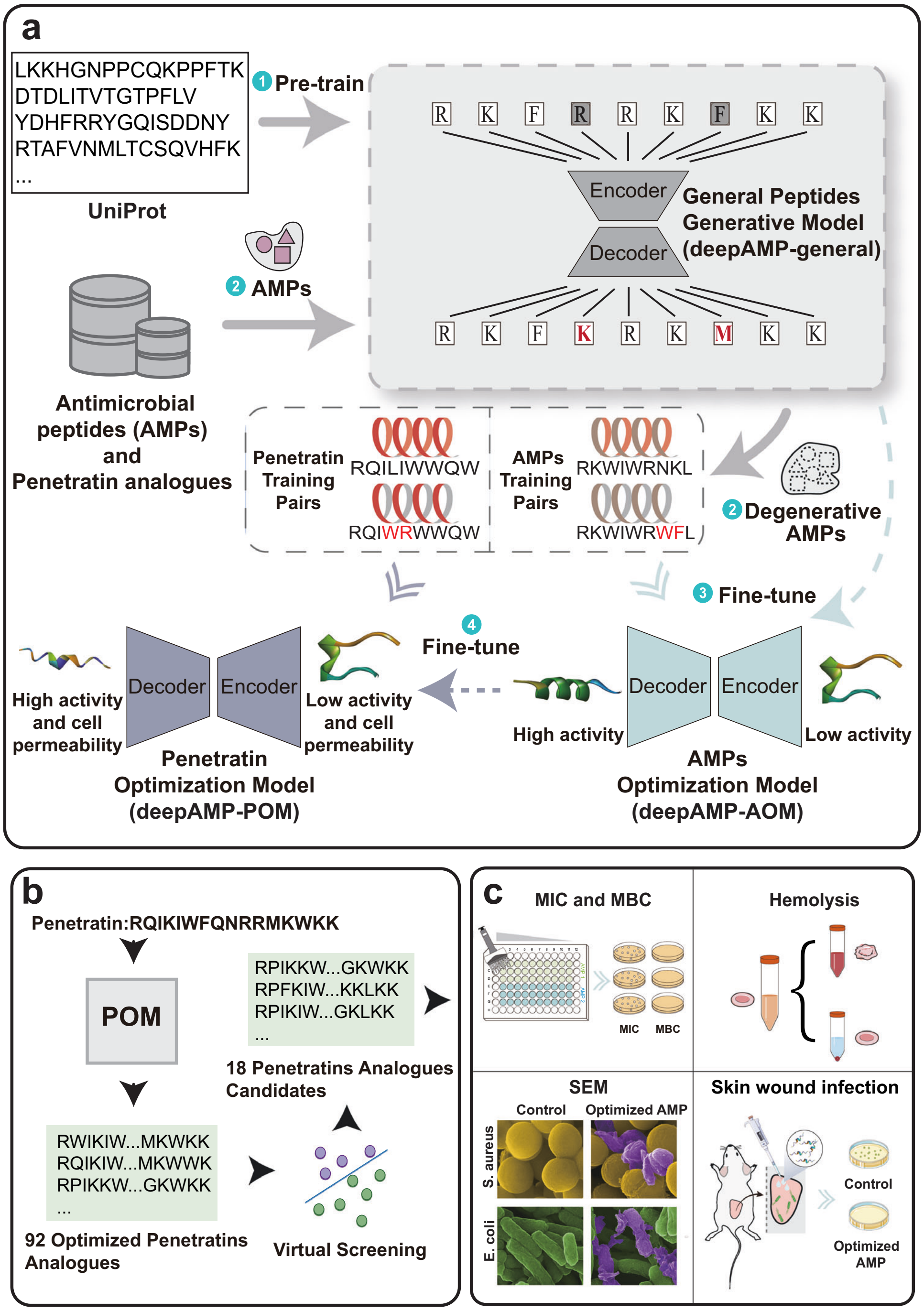
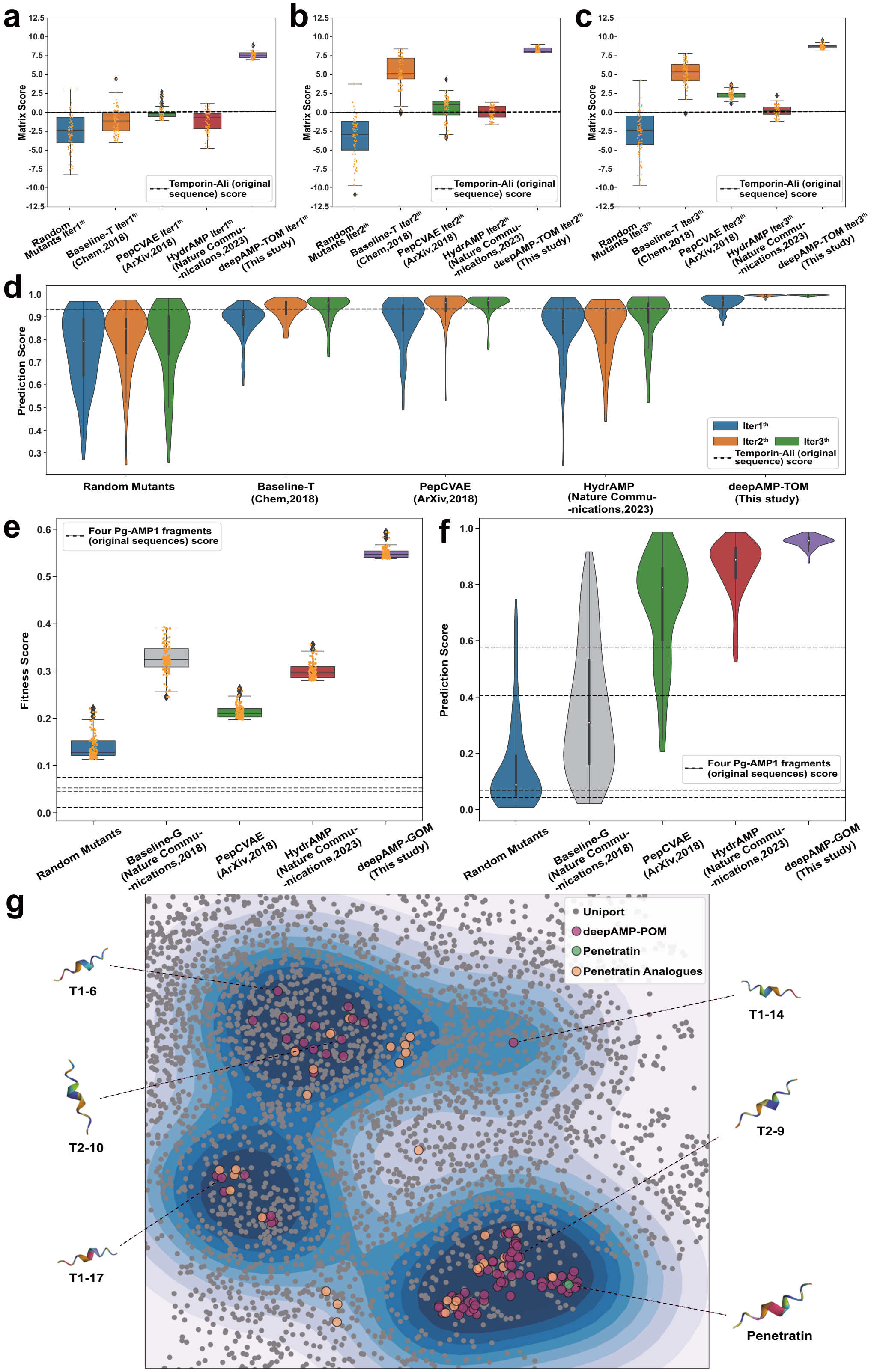


Fig. 1 | The workflow of penetratin optimization. a First, we pretrained our model with UniProt data. Secondly, we constructed Amps Training Pairs and Penetratin Training Pairs datasets (*cf.* Methods) by a degenerative workflow, and feed these to finetune the pretrained model. The deepAMP-POM was used to optimize penetratin model, and the **b** deepAMP-predict were used to rank these candidates. First round18 sequences were selected from ranked candidates for **c** in vitro and in vivo experimental validation.

to obtain deepAMP-GOM. This model was specifically applied to the optimization task for guavanin antimicrobial peptides.

Our optimization process was conducted over five rounds of iterations, selecting the top 100 candidates with the highest fitness scores in each iteration (Supplementary Fig. 1). Specifically, deepAMP-GOM surpassed the fitness scores of Baseline-G as early as the third iteration (Supplementary Fig. 1a). Our method improved the top 100 scoring candidates over five rounds of iterations, performing better than existing methods and achieving the highest score of 0.594 (Fig. 2e). These candidates were then evaluated using the deepAMP-predict scoring (Fig. 2f). The density of the scores for deepAMP-GOM also indicate our model's effectiveness in identifying potent antimicrobial peptides with greater accuracy and reliability compared to other computational approaches (Supplementary Methods and Results).

After fine-tuning deepAMP-AOM using the penetratin pairs, we generated 92 penetratin-optimized sequences by deepAMP-POM. We explored the distribution of penetratin sequences in the chemical space after deepAMP-POM optimization (termed the penetratin optimization model). We utilized the Uniform Manifold Approximation and Projection[24] (UMAP) release to downscale the model output features of the sequences. As shown in Fig. 2g, the optimized sequences



are similar to the chemical spatial distribution of penetratin and its analogues. The distribution of deepAMP-generated candidate AMPs in 3 parts shows the diversity of the model generation (the location of the distribution is determined by the one-dimensional sequence structure). Meanwhile, we utilized the mutation site as an indicator to assess the top-predicted AMPs (T1-1 to T2-11). From Supplementary Data 1, we found that our 28 top-predicted AMPs had mutation distance more than 1 from the training set, and only one predicted AMPs has mutation distance equal to 1. The average of the mutation distances between the top predicted AMPs and AMPs from the training set is 4, indicating the chemical novelty of top-predicted AMPs compared to the training set. The fine-tuned deepAMP-POM is able to identify potential AMPs with antimicrobial ability and cell permeability (Supplementary Methods and Results).

Fig. 2 | Performance of deepAMP. a Temporin-Ali optimised performance in matrix score, comparing deepAMP-TOM (purple), Baseline-T (orange), Random mutants (blue), PepCVAE (green) and HydrAMP (red) in one iteration (n = 91 candidate sequences), **b** the second iteration (n = 91 candidate sequences) and **c** the third iteration (n = 91 candidate sequences). The dashed line represents the matrix score for Temporin-Ali (matrix score is −0.17). Boxplots show the median (center line), and 1st and 3rd quartiles (Q1 and Q3, respectively). The whiskers (error bars) indicate the range of the data, defined as the range between Q1-1.5*IQR and Q3 + 1.5*IQR. **d** Temporin-Ali optimised performance in deepAMP-predict score (n = 91 candidate sequences). The dashed line represents the predict score for Temporin-Ali (predict score is 0.93). Violin plots show the median (white point), and 1st and 3rd quartiles (Q1 and Q3, respectively). The upper and lower bounds of the violin represent the minimum and maximum values of the data. **e** Pg-AMP1 optimised performance in fitness score, comparing deepAMP-GOM (purple), Random mutants (blue), PepCVAE (green) and HydrAMP (red) to the 100 candidates with the highest fitness scores in iteration (n = 100 candidate sequences). The dashed line represents the fitness score for four fragments of Pg-AMP1 (fitness score are 0.075, 0.049, 0.046, 0.012, respectively). Boxplots show the median (center line), and 1st and 3rd quartiles (Q1 and Q3, respectively). The whiskers (error bars) indicate the range of the data, defined as the range between Q1 and 1.5*IQR and Q3 + 1.5*IQR. **f** Pg-AMP1 optimized performance in deepAMP-predict score (n = 100 candidate sequences). The dashed line represents the predict score for four fragments of Pg-AMP1 (predict score are 0.57, 0.40, 0.07, 0.04, respectively). Violin plots show the median (white point), and 1st and 3rd quartiles (Q1 and Q3, respectively). The upper and lower bounds of the violin represent the minimum and maximum values of the data. (**g**) The visualization of sequences under the UMAP 2-dimensional space (Source data are provided as a Source Data file).

Experimental validation of candidate AMPs

A library of 92 peptide sequences generated by substituting amino acid residues of the penetratin are provided in the Supplementary Data 2. The top 18 AMPs, named T1-1 to T1-18, with high predicted scores based on the deepAMP-predict model, were measured by minimum inhibitory concentration (MIC; lower value is better) and minimum bactericidal concentration (MBC; lower value is better). We tested MIC and MBC against Gram-positive *Staphylococcus aureus* (*S. aureus*), Methicillin-resistant *Staphylococcus aureus* (*MRSA*), Gram-negative *Escherichia coli* (*E. coli*), *Klebsiella pneumoniae* (*K. pneumoniae*), and *Pseudomonas aeruginosa* (*P. aeruginosa*) (see Methods). We found that penetratin exhibited greater antibacterial activity against Gram-negative *E. coli* than Gram-positive *S. aureus*. For *E. coli*, a total of 16 deepAMP identified AMPs had MIC < 62.5 μg/mL, which are stronger than penetratin. In particular, 4 AMPs (T1-2, T1-7, T1-14, and T1-15) reached MIC of 3.91 μg/mL, which are comparable to known FDA-approved antibiotic levofloxacin (2.93 μg/mL). Also, a total of 13 AMPs reached <125 μg/mL in their MIC against *P. aeruginosa*, revealing stronger antimicrobial activity than penetratin. Notably, MIC values of T1-2 and T1-5 are 11.72 μg/mL, which was equivalent to levofloxacin (MIC = 7.8 μg/mL) (Table 1). *K. pneumoniae* is an opportunistic pathogen that mostly affects people with weak immune systems and causes nosocomial infections[25]. We observed that T1-2, T1-5, and T1-6 showed potent in vitro activity against *K. pneumoniae* with a MIC of 15.63 μg/mL, whereas penetratin showed very weak activity (MIC > 1000 μg/mL). In addition, T1-2 and T1-5 are the two best AMPs with 11.72 μg/mL in suppressing of *S. aureus* effectively.

To further assess the ability of the selected AMPs against multidrug-resistant (MDR) strains, we performed growth inhibition assays against difficult-to-treat resistant Gram-positive bacteria, *MRSA* strain. We found that antibiotic-resistant bacteria are sensitive to antimicrobial peptides. 88.8% (16/18) of the peptides that we tested exhibited greater antibacterial activity than penetratin. Notably, T1-4 and T1-5 were the most effective, with an MIC of 15.63 μg/mL, which was 64-fold lower than that of the penetratin (>1000 μg/mL). In particular, deepAMP-predicted AMPs inhibit both Gram-positive and Gram-negative bacteria during MIC assays, while vancomycin (the last line of medication) only targets Gram-positive bacteria. To determine the bactericidal potency of deepAMP-predicted AMPs, the MBC assay was applied. As seen in Fig. 3c (Supplementary Data 3), the observed MBC of multiple deepAMP-predicted AMPs (e.g.T1-4 against *E. coli*, T1-11 against *S. aureus* and *MRSA* were identical to their MIC, indicating that they were endowed with minimal tolerance (MBC/MIC) as well as with both bacteriostatic and bactericidal properties. Additionally, the detailed CFU count under MIC concentration are shown in Supplementary Fig. 2. We found that T1-8 and T1-11 could fully eliminate bacteria at their MIC concentrations, preventing them further proliferation. However, bacteria continue to proliferation after stopping treatment for several other candidate AMPs (T1-2 and T1-7). Further improving anti-bacterial resistance of candidate AMPs using the updated deepAMP are highly warranted.

We next turned to conduct the second generation of optimization to identify more potent AMPs using our deepAMP-POM model. We generated a library of 11 deepAMP-predicted candidate peptides (termed T2-1 to T2-11) by substituting amino acid residues of T1-5, the best candidate AMP with potent in vitro activity against both Gram-negative and Gram-positive, including multidrug-resistance. We found the effective AMP proportion in the second-generation was higher than the first generations (Table 1). Notably, the second-generation AMPs showed more potent MIC than penetratin against *E. coli* and *P. aeruginosa* (10 AMPs among 11 [90.9%]), and *K. pneumoniae*, *S. aureus*, and *MRSA* (all 11 AMPs with 100%) (Fig. 3e). The average MIC maximum value decreases 7.9 folds as compared with the first generations (343.75 /43.32, against *K. pneumoniae*). In particular, among 11 AMPs, T2-9 is the most potent peptide with the MIC value of 10.6 ~ 64-fold lower than that of the penetratin (2.93 μg/mL for *E. coli*, 15.63 μg/mL for *K. pneumoniae*, 11.72 μg/mL for *P. aeruginosa*, 15.63 μg/mL for *S. aureus* and 23.44 μg/mL for *MRSA*). In summary, combining deepAMP prediction and systematic experimental validation, we identified highly potent candidate AMPs (such as T2-9) for follow-up studies.

To evaluate the safety of AMPs in our study, we conducted hemolytic activity test and MTT (3-[4,5-dimethylthiazol-2-yl]-2,5 diphenyl tetrazolium bromide) assay. As shown in Fig. 3d, our deepAMP framework identified candidate AMPs exhibited wide variability in red blood cell (RBC) hemolysis. Among all candidates, T1-6, T1-14, T2-9 and T2-10 not only exhibited high antibacterial activity, but also possessed potent biological compatibility with low RBC hemolysis (<20%). And the concentration for 50% of maximal effect (EC_{50}) of Penetratin, T1-2, T1-5 and T2-10 was more than 100 μM (230 μg/mL, Supplementary Fig. 3a and 3b). Furthermore, Penetratin, T1-2, T1-5 and T2-10 showed low cytotoxicity to both Normal Human Dermal Fibroblasts (NHDF) cells and stem cells from human exfoliated deciduous teeth (SHED) cells. According to Supplementary Fig. 3c and 3d, the IC_{50} of these peptides exceeds 50 μg/mL, which is more than their MIC values. Taken together, the deepAMP identified candidate AMPs with potent anti-bacterial activity and low toxicity.

Structural activity relationship analysis of deepAMP-designed AMPs

To investigate the structure activity relationship (SAR) of candidate AMPs predicted by deepAMP and their biological activities (such as MIC), we calculated a series of physicochemical properties[26], such as hydrophobicity, hydrophobic moment and Boman index. The results showed moderate correlations between the calculated parameter values and MIC. For example, the correlation between the hydrophobic moment and MIC value was 0.47 towards *K.pneumoniae* (Supplementary Figs. 4–8). We found that among peptides with lower hemolysis most possessed lower hydrophobic fraction (under 0.4) and

Table 1 | The sequence of all the peptides and antibiotics, as well as their minimum inhibitory concentration values

		Antimicrobial activity (MIC in μg/ml)									
Peptide	**Sequence**	***E. coli***	**±SEM**	***S. aureus***	**±SEM**	***K. pneumoniae***	**±SEM**	***P. aeruginosa***	**±SEM**	**MRSA**	**±SEM**
Penetratin	RQIKIWFQNRRMKWKK	62.50	0.00	250.00	0.00	1000.00	0.00	125.00	0.00	1000.00	0.00
T1-1	RWIKIWFQIRRWKWKK	11.72	5.52	15.63	0.00	31.25	0.00	31.25	0.00	23.44	11.05
T1-2	RQIKIWFQIRRWKWKK	3.91	0.00	11.72	5.52	15.63	0.00	11.72	5.52	31.25	0.00
T1-3	RQIKIWFQNKKWKWKK	62.50	0.00	250.00	0.00	1000.00	0.00	125.00	0.00	1000.00	0.00
T1-4	RWIKIWFQIRRWKNKK	7.81	0.00	15.63	0.00	23.44	11.05	31.25	0.00	15.63	0.00
T1-5	RQIKIWFQWRRWKWKK	5.86	2.76	11.72	5.52	15.63	0.00	11.72	5.52	15.63	0.00
T1-6	RQIKIWLQIRRWKNKK	7.81	0.00	31.25	0.00	15.63	0.00	15.63	0.00	62.50	0.00
T1-7	RWIKIQLQIRRWKNKK	3.91	0.00	31.25	0.00	62.50	0.00	62.50	0.00	93.75	44.19
T1-8	RWIKIWFQNRRWKWKK	31.25	0.00	15.63	0.00	375.00	176.78	93.75	44.19	62.50	0.00
T1-9	RWIKIWFQWRRWKWKK	15.63	0.00	31.25	0.00	62.50	0.00	62.50	0.00	46.88	22.10
T1-10	RQIKIWFQNRKWKWKK	31.25	0.00	250.00	0.00	500.00	0.00	62.50	0.00	250.00	0.00
T1-11	RQIKIWFQNKRWKWKK	31.25	0.00	250.00	0.00	500.00	0.00	125.00	0.00	500.00	0.00
T1-12	RQIKIWFQNRRWKWKK	15.63	0.00	125.00	0.00	500.00	0.00	93.75	44.19	250.00	0.00
T1-13	RWIKIWFQNRRWKNKK	15.63	0.00	23.44	11.05	500.00	0.00	187.50	88.39	125.00	0.00
T1-14	RQIKIWFQIRRWKNKK	3.91	0.00	31.25	0.00	23.44	11.05	31.25	0.00	62.50	0.00
T1-15	RWIKIQFQIRRWKNKK	3.91	0.00	15.63	0.00	62.50	0.00	46.88	22.10	62.50	0.00
T1-16	RQIKIWFQNRRWRWKK	15.63	0.00	62.50	0.00	500.00	0.00	93.75	44.19	125.00	0.00
T1-17	RWIKIQLQNRRWKNKK	62.50	0.00	250.00	0.00	1000.00	0.00	500.00	0.00	1000.00	0.00
T1-18	RWIKIQFQNRRWKNKK	31.25	0.00	125.00	0.00	1000.00	0.00	375.00	176.78	500.00	0.00
T2-1	RWIKIWFQWRKIRWKK	7.81	0.00	23.44	11.05	31.25	0.00	31.25	0.00	46.88	22.10
T2-2	RQIKIWFIWRKWRWKK	15.63	0.00	31.25	0.00	23.44	11.05	62.50	0.00	62.50	0.00
T2-3	RWIKIWFQIRRWKWRK	15.63	0.00	23.44	11.05	31.25	0.00	62.50	0.00	31.25	0.00
T2-4	RWIKIWFQIRRWRWKK	15.63	0.00	31.25	0.00	46.88	22.10	62.50	0.00	31.25	0.00
T2-5	KWIKIWFQWRRWRWKK	7.81	0.00	15.63	0.00	31.25	0.00	46.88	22.10	31.25	0.00
T2-6	KWIKIWFQWRRWKWKR	15.63	0.00	23.44	11.05	31.25	0.00	31.25	0.00	31.25	0.00
T2-7	KWIKIWFIWRRIKWRK	15.63	0.00	15.63	0.00	46.88	22.10	46.88	22.10	31.25	0.00
T2-8	KQIKIWFIWRRIKIKK	7.81	0.00	31.25	0.00	62.50	0.00	31.25	0.00	62.50	0.00
T2-9	RQIKIWFQWKRIKWRR	2.93	1.38	15.63	0.00	15.63	0.00	11.72	5.52	23.44	11.05
T2-10	RQIRIWFQWKRWKWKR	5.86	2.76	23.44	11.05	31.25	0.00	23.44	11.05	31.25	0.00
T2-11	RWIKIWFIWRRWRWRK	62.50	0.00	62.50	0.00	125.00	0.00	187.50	88.39	125.00	0.00
Vancomycin		125.00	0.00	0.98	0.00	1000.00	0.00	1000.00	0.00	0.98	0.00
Levofloxacin		2.93	1.38	0.09	0.04	0.06	0.00	7.81	0.00	0.48	0.00

hydrophobic moment (under 0.5). Both hydrophobic fraction and hydrophobic moment rose to more than 0.5 in the second generation, indicating that these two properties may contribute to the hemolysis (Supplementary Data 4). Boman indexes were calculated to predict the physical binding of peptides. We found that except for T2-7 and T2-8, the Boman indexes of deepAMP-predicted AMPs were higher than 2.48, suggesting strong binding potential of our designed candidate AMPs[27] (Supplementary Data 4).

Hydrophobic profile was related to the amphiphilicity of α-helical peptides[28]. The increased hydrophobic moment indicated that the deepAMP-predicted AMPs improved the spatial amphiphilicity (designed AMPs with average value of 0.50 vs. penetratin 0.37) (Supplementary Data 4). To inspect that deepAMP identified AMPs adopt α-helix structures, we performed de novo structural reconstruction for T1-2 using Discovery Studio (see Methods) and found the similar helical structure as penetratin (Supplementary Fig. 9a). In terms of the peptide sequence, the helical wheels showed that they had the same positively charged amino acid residues (Supplementary Fig. 9b). The peptides were further experimentally characterized using CD spectroscopy. Both T1-2 and penetratin formed an α-helix in DOPC/DOPG large unilamellar vesicles (LUVs) (Supplementary Fig. 9c). From the CD spectra, the α-helicity of T1-2 is stronger than that of penetratin, suggesting that its stronger hydrophobic profile may allow better disruption of the bacterial membrane (Supplementary Data 4).

Mechanism-of-Action of AMPs designed by deepAMP

Previous studies have shown that AMP inactivates pathogenic microbes primarily by interfering with and damaging their cell membranes[17]. Based on CD spectroscopy of T1-2, we first investigated AMPs in the presence of the membrane using all-atom molecular dynamics simulation. Our findings showed that the peptide was embedded into the cell membrane (Supplementary Fig. 9d–e), with its hydrophilic region facing the lipid headgroup, and its hydrophobic portion facing the membrane. The peptide maintained its helical conformation throughout the 500 ns simulation, with its distance to the membrane center ranging between 2.2-2.6 nm (Supplementary Fig. 9f). Our analysis revealed key interactions between the peptide and the lipid molecules, including aromatic amino acids (phenylalanine and tryptophan) interacting with the lipid glycerol group and cholesterol, and the positively charged arginine and lysine interacting with the phosphatidylcholine headgroup. As a result of peptide insertion, the lipid acyl chains in the vicinity of the peptide experienced significant disorder (Supplementary Fig. 9g), which could lead to excessive pressure, membrane rupture, and ultimately cell lysis.

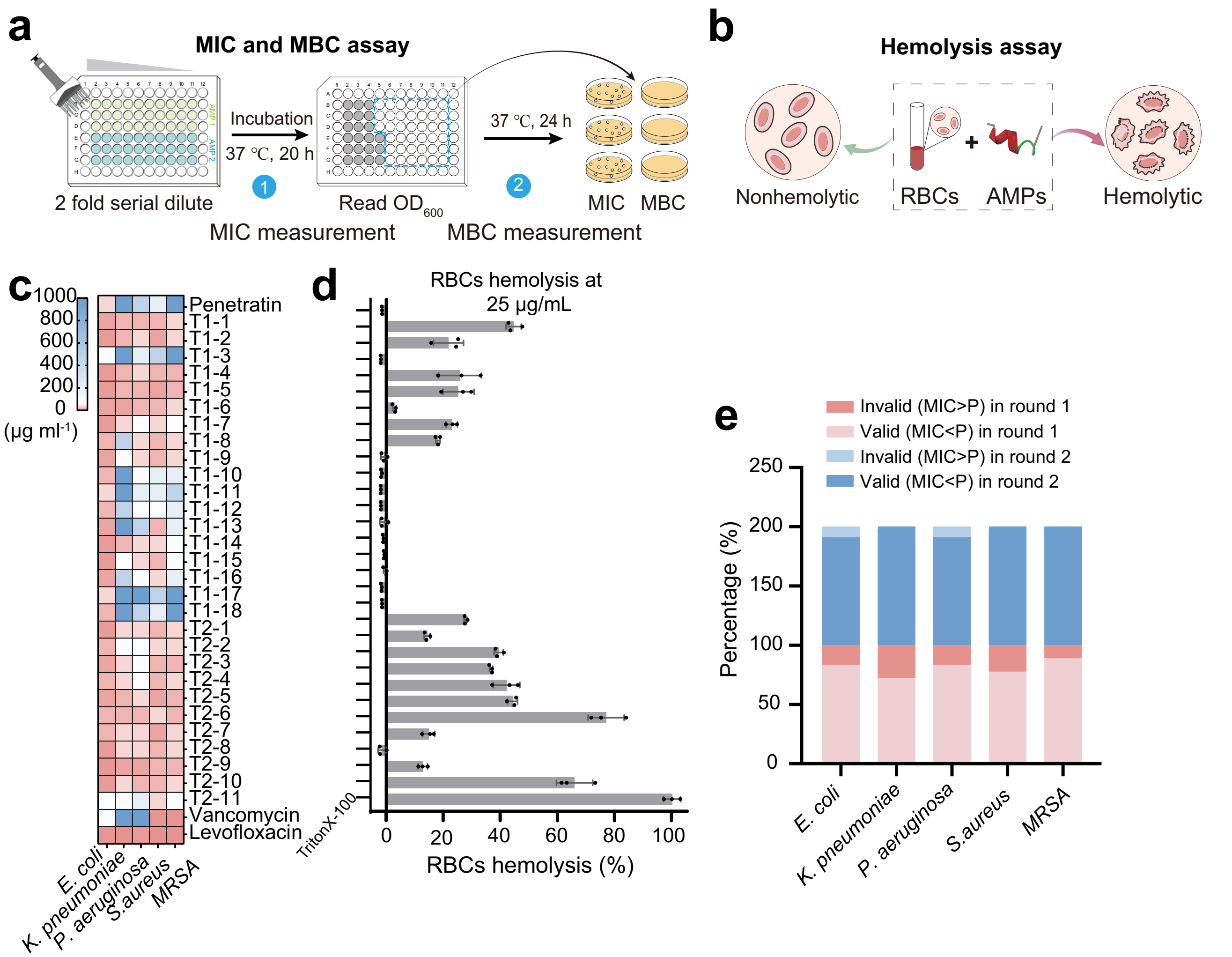


Fig. 3 | Antimicrobial and hemolytic activities of candidate AMPs. a Schematic illustration of the method of MIC and MBC test. (**b**) Schematic illustration of the method of hemolysis test. **c** The MBC value of candidate AMPs, $n = 3$. **d** The hemolytic activities of candidate AMPs, $n = 3$. **e** The percentage of peptide which MIC is lower than Penetratin in total 30 peptides (Source data are provided as a Source Data file).

To test the mechanism-of-action of deepAMP-predicted candidate AMPs further, we conducted four complementary experiments: (i) propidium Iodide (PI) staining detected by confocal and flow cytometry, (ii) scanning electron microscopy (SEM), (iii) membrane depolarization assay, and (iv) permeabilization of the outer membrane assay. We first used PI (only stains dead cells) and Hoechst (stains both live and dead cells) to conduct live and dead assays under AMPs treatment. We found that both *E. coli* and *S. aureus* emitted red fluorescence after treatment with T1-2 and T2-10, suggesting that the AMPs induce pore formation on the bacterial membrane and enable the PI dye to enter the bacteria (Fig. 4a). Similar fluorescence was also observed in *K. pneumoniae*, *P. aeruginosa*, and *MRSA* (Supplementary Fig. 10). The elevated PI intensity in flow cytometry assay provided additional evidence that AMPs caused damage to the bacterial membranes (Fig. 4b and Supplementary Fig. 11).

The effects of treatment on bacteria with AMPs were further determined by scanning electron microscopy (SEM). As shown in Supplementary Fig. 12, untreated *S. aureus* showed a spherical shape and smooth surface. While after AMP treatment (T1-2, T1-5, T2-9, and T2-10), the bacteria had a significant influence on the morphology of the bacterial cell surface with membrane damage and wrinkled surface fragments. Similar observations were obtained for *E. coli* and *P. aeruginosa*. We next assessed the ability of AMPs to damage and depolarize the bacterial cytoplasmic membrane. 3,3′-dipropylthiadicarbocyanine iodide ($DiSC_3(5)$) is a potentiometric fluoroprobe. Upon changes in the cytoplasmic membrane transmembrane potential, the fluorophore transfers from the cytoplasmic membrane to the outer environment, generating fluorescence. We found the $DiSC_3(5)$ fluorescence under AMP (T1-2, T1-5, and T2-10) treatment was 2-4 times that of untreated samples. In particular, the fluorescence of T2-10 was 1.9 times higher than polymyxin B, a positive control (Fig. 5b). This result illustrates that AMPs may inhibit bacteria by breaking the cytoplasmic membrane. A fluorescent probe 1-N-phenylnaphthylamine (NPN) uptake assay was further performed to assess whether AMPs permeabilized the outer membrane. NPN does not penetrate the bacterial outer membrane unless it is broken. When bacteria were treated with deepAMP identified AMPs, the fluorescence was twice as high as untreated, which is comparable to the positive control of polymyxin. These observations suggest that AMPs exhibited high potency to permeabilize the outer membrane (Fig. 5c).

In summary, these comprehensive assays suggested that deepAMP-designed candidate AMPs not only influence the cytoplasmic membrane of bacteria but also exert their inhibitory effects by permeabilizing the outer membrane, indicating potency to overcome antibacterial resistance by targeting bacterial membranes.

deepAMP-design AMPs reduce antibiotic resistances

We next turned to test whether deepAMP identified AMPs stimulate bacterial resistance. We performed a resistance-acquisition test on *S.*

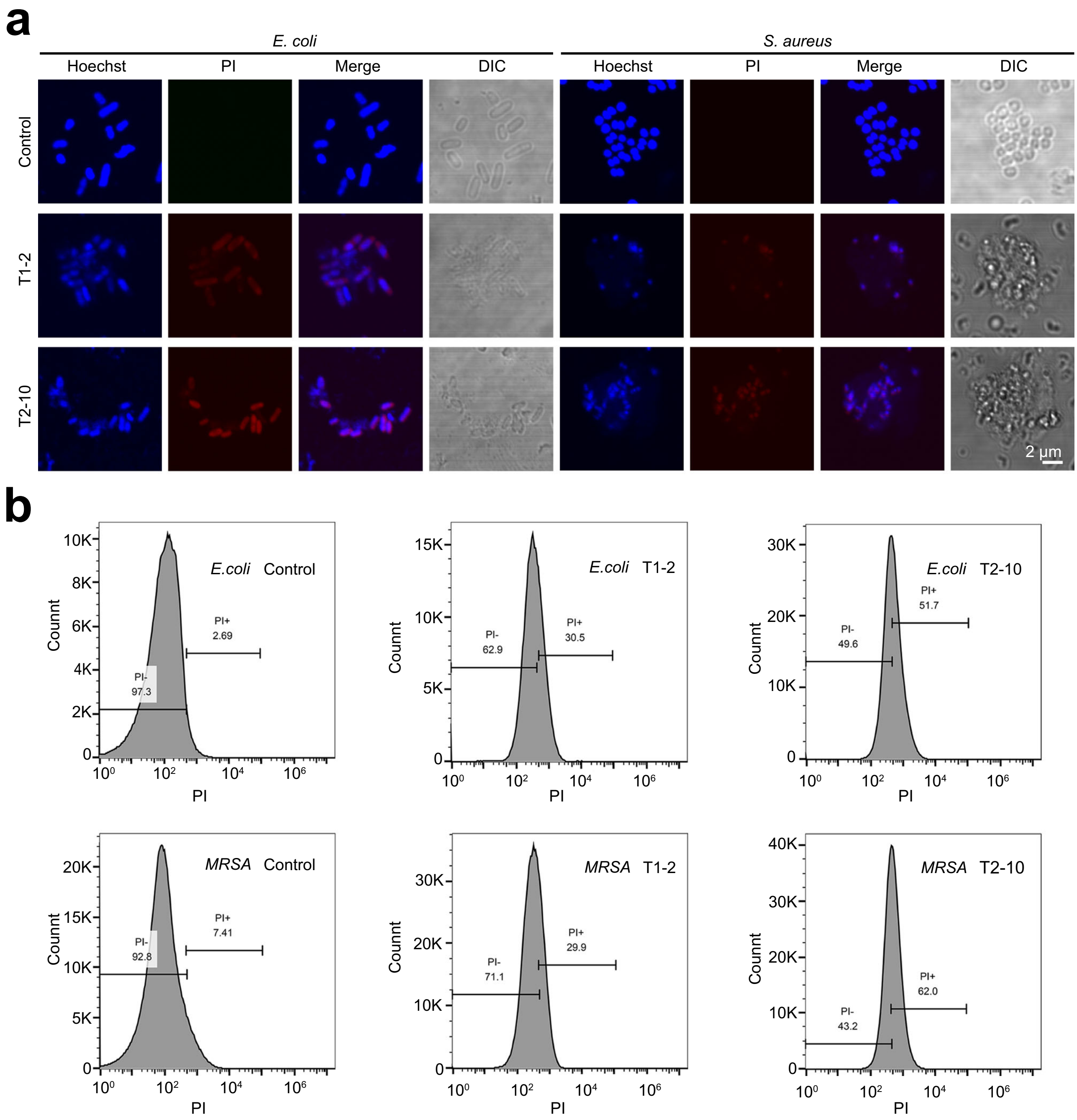


Fig. 4 | The effect of candidate AMPs on the bacterial membrane. a Live-dead staining of *E. coli* and *S. aureus* treated with T1-2 and T2-10. Each experiment was repeated independently at least three times with similar results and the gating strategies show as Supplementary Fig. 11a. **b** Membrane permeability assay using *E. coli*, *MRSA*, and propidium iodide (PI), show bacterial cell membrane disruption by the addition of T1-2 and T2-10 (1 × MIC). Compound treatment time was 5 h for *E. coli* and *MRSA* (Source data are provided as a Source Data file).

aureus in the presence of ciprofloxacin (an FDA-approved quinolone antibiotics), T1-2, T1-5, and T2-10. We found that three peptides T1-2, T1-5, and T2-10 did not induce resistance after 25 passages (Fig. 5d), whereas *S. aureus* began to develop resistance to the antibiotic ciprofloxacin after only 10 passages. Bacterial biofilm helps antibiotics tolerance[29], thus, we next test the ability of our AMPs to inhibit biofilm formation using crystal violet staining. As shown in Fig. 5e and Supplementary Fig. 13, we found that three deepAMP-designed AMPs (T1-2, T1-5, and T2-10) were effective in preventing biofilm formation towards both *S. aureus* and *E. coli*. Altogether, these experimental observations indicate that T1-2, T1-5, and T2-10 have therapeutic potential for treating resistant strains and hence demand further experimental and clinical investigation in the future.

AMPs are effective against skin wound infection in vivo

Dermis wound infection is one of the health problems caused and exacerbated by the invasion of pathogenic organisms[30]. We harnessed a mouse wound infection model with *P. aeruginosa*, one of the most common causes of chronic wound infections, to evaluate the in vivo efficacy of T1-2, T1-5 and T2-10. As shown in Fig. 6b and Supplementary Fig. 14a, 107 CFU bacteria were decreased to about 0 CFU after 24 h treatment of AMPs, achieving the curative effect of Levofloxacin

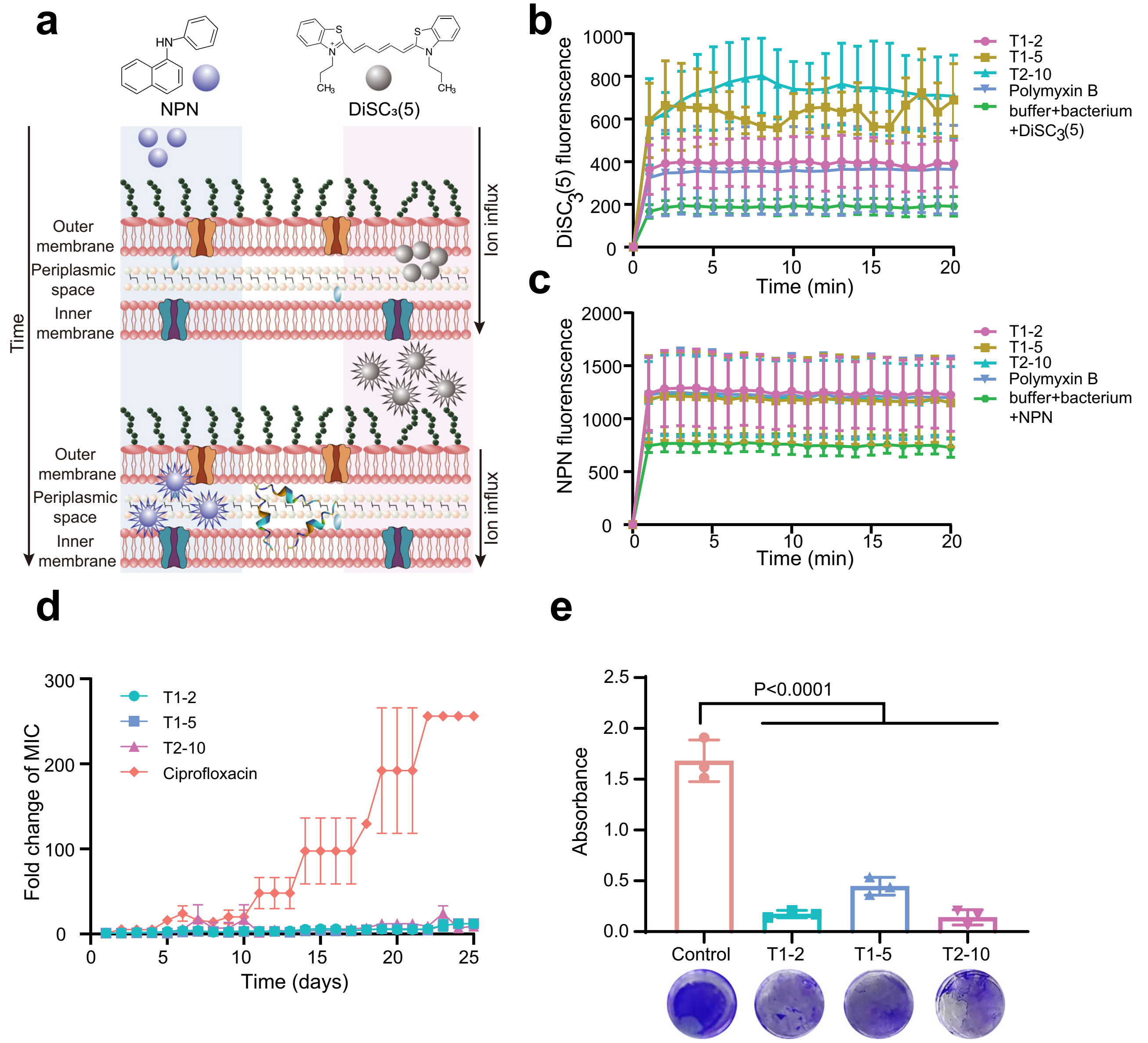


Fig. 5 | Mechanism of action studies and resistance development of candidate peptides. a Schematic showing increased fluorescence resulting from membrane instability (left panel, purple) and depolarization (right panel, grey) caused by the peptides at their MIC over time. **b** $DiSC_3(5)$ assays show the effect of peptides on the permeabilization of the outer membrane of *E. coli*, *n* = 3. **c** NPN assays show the effect of peptides on the permeabilization of the outer membrane of *E. coli*, *n* = 3. **d** Evolution of resistance to T1-2, T1-5, T2-10, or ciprofloxacin in *S. aureus* after 25 days of passaging in liquid MH media. Cells were passaged every 24 h, *n* = 4. **e** Crystal violet staining image and its corresponding absorbance for integrated *S. aureus* biofilm incubated with the peptides for 72 h, *n* = 3. The biofilm under without incubation with the peptide was used as the control. All data are mean ± s.d. and were analysed by one-way ANOVA with Dunnett's multiple comparison test. (Source data are provided as a Source Data file).

(0 CFU). In contrast, more than 10^5 CFU bacteria were detected on PBS treated skin surface. A qualitative test H&E staining was used to reflect the antimicrobial effects after 3 d treatment. Despite it wasn't quantified, it is evident that mice in the PBS group exhibited a higher degree of neutrophil infiltration compared to the three other groups treated with T1-2, T1-5, and T2-10, respectively (as indicated by the green arrows in Supplementary Fig. 14b). In addition, we did not observe organ-specific toxicity in mice after treatments in the tissue immunohistology section images (Supplementary Fig. 15). Blood test results showed no obvious changes among different treatment groups (Supplementary Fig. 16). In summary, these mouse model observations reveal that the deepAMP-identified AMPs display potent in vivo effects to treat *P. aeruginosa* skin infection and exhibit good biocompatibility.

Discussion

Antibiotic resistance crisis is sweeping the world and the steadily declining productivity of new clinical antibiotics due to the high risk of early discovery and poor return on investment is exacerbating this problem. AMPs as potential alternatives to antibiotics with low risk of drug resistance has attracted growing attention. In this study, we created a language model-based antimicrobial peptide discovery framework (termed deepAMP) and experimentally validated the candidate peptides for broad-spectrum antimicrobial action and treating resistant bacteria using both in vitro and in vivo models.

Several computational approaches[15,17] have been developed for virtual screening for AMP discovery. These methods are limited by the small data that only identify potential AMPs in the restricted chemical

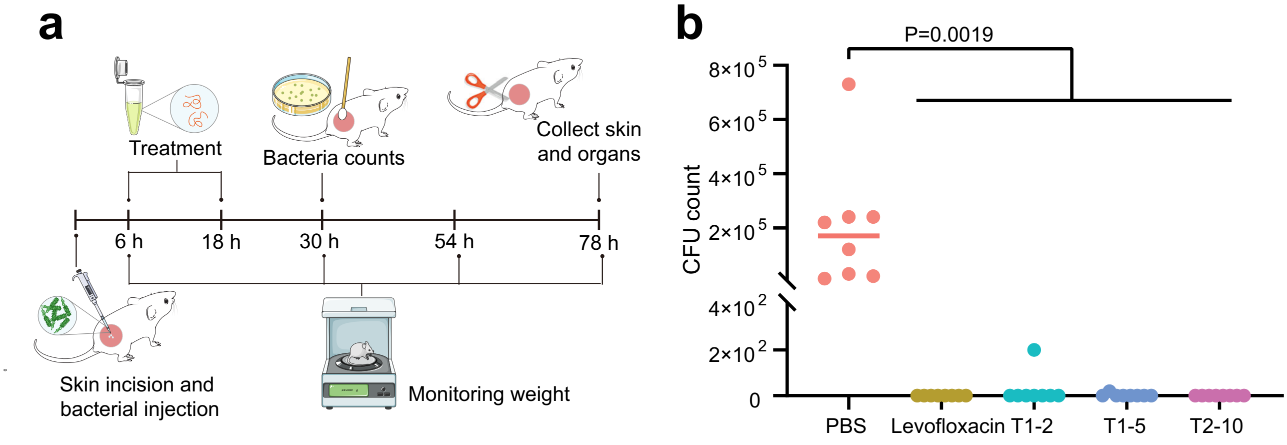


Fig. 6 | Anti-infective activity of candidate peptides in vivo. a Schematic of the mice excision wound model used to assess the anti-infective activity of AMPs. **b** Significant reduction of the number of bacteria on the infected areas by Levofloxacin, T1-2, T1-5, and T2-10 (10 mg/kg) in the mice excision wound model study, *n* = 8. All data are mean ± s.d. and were analysed by one-way ANOVA with Tukey's multiple comparison test. (Source data are provided as a Source Data file).

space. Although generative models[18,19,23] were proposed for de novo AMPs design, unconstrained generation results in difficult synthesis and high experimental cost of the generated AMPs. Our deepAMP framework learned a universal representation of peptide sequences in an unsupervised manner from a large and varied sequence database of peptides, ensuring the diversity of the generated AMP candidates. Existing methods might not be able to effectively capture high-dimensional features leading to poor optimization results and low efficiency[16,21]. DeepAMP was fine-tuned on few-shot data to learn the key to AMP switching between low and high antimicrobial activities. Both in vitro and in vivo assays confirmed that the deepAMP identified candidate AMPs are tens of times more potent than the original sequence, and less susceptible to antibacterial drug resistance.

We acknowledged several potential limitations. One-dimensional peptide sequence generation implemented in the current deepAMP framework cannot capture three-dimensional (3D) structural information. Integrating 3D conformational informations derived from existing approaches, such as alphFold3[31], may further improve performance of deepAMP. The peptide language model is also a black-box model by lacking interpretability, which hinders feature characterization of bioactive peptides. In the future, we will develop multimodal peptide generation methods to capture both sequence and 3D structural information of peptides and improve model interpretability. While the initial experimental validation was limited to 29 candidates, this represents a relatively small size. More experimental validations for a large number of candidate AMPs are warranted in the future. Furthermore, given the potent antibacterial effects observed with the three lead AMPs (T1-2, T1-5, and T2-10) in the wound mouse model, more in vivo observations of these promising AMPs using additional animal models, such as those for deep thigh infection and pneumonia mouse models, are warranted as well.

In summary, our study demonstrates a practical application of using large-scale language models for the discovery of AMPs with high potent antimicrobial activities and low antibiotic resistance. As the proof of concept, we demonstrated that more than 90% of the designed candidate AMPs showed a better MIC (minimal inhibitory concentration) than penetratin[8–11]. AMP T2-9 showed the highest antimicrobial activity against five types of bacteria, with 10 ~ 64 times of higher potency than penetratin. Confocal microscopy and flow cytometry assays showed strong membrane permeability of designed AMPs and three candidates (T1-2, T1-5 and T2-10) displayed a low propensity to induce resistance in *S. aureus* compared to ciprofloxacin (an FDA-approved antibiotic). Mechanistic observations revealed that designed AMPs selectively disrupt cell membrane of bacteria and overcomes antibiotic resistance. In summary, our deepAMP framework offer a powerful deep generative model for discovery of potent AMPs and other types of therapeutic biologics if broadly applied.

Methods

Model pre-training and fine-tuning

We follow the idea of pre-training and fine-tuning (network parameters in Supplementary Tables 1, 2), which has been shown to work well on tasks with small datasets. To better fit the task requirements, we divided the model into 3 parts, (1) the peptide generative model (deepAMP-general), (2) the antimicrobial peptide optimization model (deepAMP-AOM), and (3) the penetratin optimization model (deepAMP-POM). The model details are described in the following.

Data sets of AMPs

Initially, we aimed to obtain a training dataset de novo that was inexpensive, chemically diverse, and did not require sophisticated laboratory resources. This would allow for the development of a robust model with which new AMPs could be predicted and optimized, without the practical hurdles that can be associated with large-scale antibiotic screening efforts.

We trained the three parts of the model separately with different datasets (Supplementary Table 3). (1) We constructed a pre-trained peptide dataset using the database Uniprot, which contains a large amount of protein and peptide data from different organisms and different sources, thus providing a diverse data resource for peptide research. From this, data with sequence lengths as close to peptides as possible (length range 10 ~ 50) are filtered, including 300,000 peptide sequences. (2) In order to meet the demand for antimicrobial peptide optimization, the model needs to use paired data. However, the optimized peptides in the available data differ significantly from the original peptides, in addition, there are rarely paired data samples with equivalent length and sequence similarity. For this reason, we designed an antimicrobial peptide sequence degradation method to degrade the high-activity samples into low-activity samples. We selected 321 antimicrobial peptides with relatively high activity (for E. coli MIC below 2.5) for degradation in the GRAMPA[19] dataset, where each sequence was randomly masked with up to 30% of the length of the site and was sampled 100 times. Finally, we generated 24,928 pairs containing <low activity, high activity> antimicrobial peptide optimization data set. (3) Since penetratin has been shown to have good cell permeability and average antimicrobial activity, our ultimate goal is to optimize penetratin to achieve high activity antimicrobial effect. Nevertheless, there is currently no constructed dataset for penetratin.

Here we use a data mining approach to collect data on penetratin, which includes both the original and the penetratin-optimised analogs. The data of 54 peptides against eight bacterial species were collected. After screening this part of the dataset, 29 penetratin sequences with high activity (MIC value lower than 150 μg/ml) and sequence length within 20 were selected. Then, sequence degradation was performed on these 29 sequences to generate an optimized dataset of 1009 pairs of <low activity, high activity> samples. The distribution of the predicted scores for the training set is shown in Supplementary Fig. 48.

Generalized Peptide Generative Model (deepAMP-general)

The framework of the generative model is composed of an encoder and a decoder, where we use the Transformer module based on a self-attention mechanism for the encoder and a dense network for the decoder. In the polypeptide generative model, the model uses a constructed dataset containing 300,000 peptides to train the generative model in an unsupervised manner. Since our final task requirement is to optimize the antimicrobial peptide sequences, which means only the amino acids at specific sites in the sequence need to be changed to alter the activity of the sequence, we use the masked language model (MLM) training through similar to the BERT model. Suppose the existing samples $S_i = \{s_0, s_1, \ldots, s_n\}$, we randomly mask no more than 30% of the tokens in the sequence, and the input samples of the masked model are $M_i = \{m_0, m_1, \ldots, m_n\}$, the training goal is to encode the input samples and then decode the masked token m_k into the original one by the decoder. By decoding the token m_k into the original token s_k, and the objective function of this part is:

$$L(S) = \sum_i \sum_k \log P(s_k | m_0, \ldots, m_n; \theta), \tag{1}$$

where n denotes the sequence length, k denotes the token position of the mask, $\{0 \le k \le n\}$, the number of k ranges from $(0, \lceil 30\% {*} n \rceil)$, and the conditional probability P is a generative model function with parameter θ.

The AMPs optimization model (deepAMP-AOM)

deepAMP-AOM is fine-tuned using deepAMP-general and initialized using the deepAMP-general parameters, in the fine-tuning stage, due to the difference in the amount of training sample data. To avoid overfitting the model and not learning effectively small data, it is often necessary to set a lower learning rate and increase the decay rate of weights compared to the pre-training phase. Finally, several training hyperparameters are applied to the antimicrobial peptide optimization model AOM: a batch size of 128, number of training rounds of 200, initial learning rate of 6×10^{-5}, and weight decay of 1×10^{-3}. The deepAMP-AOM model is different from the unsupervised training of deepAMP-general. In this step we utilize pairs of <low activity, high activity> training samples, the string of the input mask is the low activity sample, and the target string in decoding is the string of the mask position corresponding to the high activity sample. An optimized training set $< S_i, T_i >$ is used, where $S_i = \{s_0, s_1, \ldots, s_n\}$ is the low-activity antimicrobial peptide and $T_i = \{t_0, t_1, \ldots, t_n\}$ is the high-activity antimicrobial peptide. The model randomly masks no more than 30% of the token of low activity antimicrobial peptides, and the masked model input is $M_i = \{m_0, m_1, \ldots, m_n\}$. The goal of training is to encode the whole input sample, and after the decoder can decode the masked token m_k back to the corresponding position of high activity antimicrobial peptide token t_k. The objective function is:

$$L(S) = \sum_i \sum_k \log P(t_k | m_0, \ldots, m_n; \theta), \tag{2}$$

where n denotes the sequence length, k denotes the token position of the mask, $\{0 \le k \le n\}$, the number of k ranges from $(0, \lceil 30\% {*} n \rceil)$, and the conditional probability P is a generative model function with parameter θ.

Penetratin optimized model (deepAMP-POM)

The antimicrobial peptide contains penetratin, and with deepAMP-AOM as an intermediate layer, the model is already capable of the task of generating high-activity antimicrobial peptides, at which point we again fine-tune the parameters (Supplementary Data 5) in deepAMP-AOM to be used in the initialization of deepAMP-POM. In the training process we also classify penetratin sequences into two categories, high activity and low activity. The process and objective function are the same as deepAMP-AOM, but the data set of penetratin is small, so we need to use a smaller learning rate to learn. By adjusting the training process, we finally get a batch size of 32. The number of training rounds is 200, the initial learning rate is 1×10^{-5} and the weight decay is 1×10^{-3} .

Temporin-Ali optimized model (deepAMP-TOM). For the optimization task of Temporin-Ali, we fine-tuned deepAMP-AOM on the Temporin-Ali analog dataset to achieve better optimization. In the training process, we also classify Temporin-Ali sequences into two categories, high activity and low activity. The process and objective function are the same as deepAMP-AOM, but the data set of Temporin-Ali is small, so we need to use a smaller learning rate to learn. By adjusting the training process, we get a batch size of 32. The number of training rounds is 200, the initial learning rate is 1×10^{-5} and the weight decay is 1×10^{-3}.

Guavanin antimicrobial peptide optimized model (deepAMP-GOM)

For the optimization task of Pg-AMP1 fragment, we fine-tuned deepAMP-AOM on the guavanin antimicrobial peptide analog dataset[22] to achieve better optimization. In the training process, we also classify guavanin antimicrobial peptide sequences into two categories, high activity and low activity. The process and objective function are the same as deepAMP-AOM. By adjusting the training process, we get a batch size of 32. The number of training rounds is 200, the initial learning rate is 1×10^{-5} and the weight decay is 1×10^{-3}.

DeepAMP-based Virtual Screening

Evaluating the optimization capability of a model through virtual screening is a general method, and we used the Support Vector Machine (SVM) classifier to evaluate the antimicrobial activity of the data generated by the optimization model in advance of the wet experiment. The model is trained to distinguish the activity of antimicrobial peptides based on transforming amino acid strings into a 400-dimensional feature vector representation using the RECM feature extraction method. We collected GRAMPA[19] data as positive samples with antimicrobial activity and Uniprot peptide data as negative samples without antimicrobial activity, resulting in an antimicrobial peptide identification dataset of 6,760 positive and 6760 negative samples. Among them, 500 positive samples and 500 negative samples are assigned to the test set, and the other is the training set. The final AUC of the model is 0.96 using five-fold cross-validation, which indicates that the activity of peptide sequences can be evaluated to some extent by virtual screening.

Sequence degradation

Since optimization requires paired samples for learning, however, such data is lacking in the existing dataset. The deepAMP-general we trained is a peptide generative model, in which most of the trained peptides are inactive. Therefore, we resorted to deepAMP-general to degrade the activity of highly active sequences. Specifically, the steps of degenerating a highly active sequence are (1) random masking of the highly active sequence, the number of masks ranges from [1, length*30%). For example, for a sequence of length 10, the number of masks is at least 1 and at most 3, the mask positions are randomly combined among ten positions; (2) the sequence obtained after masking will be mapped to the hidden space of peptide by the encoder of deepAMP-general. Finally, the amino acid symbols at the mask positions will be parsed by the decoder of deepAMP-general. In this study, the process will be repeated more than 100 times for each sequence.

Fitness function

This function is used to evaluate the amphipathic α-helices of the peptide and is designed based on the ratio of Eisenberg's hydrophobic scale[32] and the sum of exponential α-helix propensity in Pace–Schols scale[33], as in Eq. 3:

$$Fitness = \frac{\sqrt{\left[\sum_{i=1}^{I} H_i \times \cos(\delta_i)\right]^2 + \left[\sum_{i=1}^{I} H_i \times \sin(\delta_i)\right]^2}}{\sum_{i=1}^{I} e^{Hx_i}}, \qquad (3)$$

where δ represents the angle between the amino acid side chains (100° for α-helix, on average); i the residue number in the position i from the sequence; H_i the ith amino acid's hydrophobicity on a hydrophobicity scale; Hx_i the ith amino acid's helix propensity in Pace–Schols scale; and I the total number of residues present in the sequence.

Peptide novelty calculation

To calculate the number of mutation sites we use the Hamming Distance, which represents the number of amino acids at different positions between two polypeptide sequences (for unequal sequences we use left-alignment for comparison). Assuming that there are two equal length peptide sequences A and B which are of length L, the number of mutation sites (Mutation Count) can be expressed as:

$$Mutation\ Count = \sum_{i=1}^{L} \delta(A_i, B_i) + |len(A) - len(B)|, \qquad (4)$$

where $\delta(A_i, B_i)$ denotes that at position $\boldsymbol{i}$. Whether the amino acids of peptide sequences A and B are different. If different, $\delta(A_i, B_i)$ is equal to 1, and if the same $\delta(A_i, B_i)$ is equal to 0, $len(A)$ denotes the length of the peptide sequence A, and $len(B)$ denotes the length of the peptide sequence B. The number of mutation sites is derived from the sum of all positions plus the difference in sequence lengths.

Peptide synthesis

Standard Fmoc solid-phase peptide synthesis was used to synthesize all peptides via a CSBio 163 peptide synthesizer. The resin-bound peptide underwent cleavage using a cocktail of TFA/thioanisole/H_2O (95:2.5:2.5) for 3 h. Collecting the filtrate and the resin mixture was washed with excess TFA. The crude peptide was obtained by concentrating and precipitating the filtrate with cold ether. Subsequently, crude peptide purification was performed through reverse phase HPLC (Elite P230A/P) using a semi-preparative C18 column. After that, all peptides were lyophilized and analyzed via analytical HPLC and ESI (+) mass.

Minimal inhibition concentration (MIC)

The bacteria S. aureus (ATCC 25904), E. coli (ATCC BAA-1025), K. pneumoniae (ATCC27736), P. aeruginosa (ATCC 47085) and MRSA (ATCC BAA-1717) were kindly donated by Prof. Xinxin Feng's group. 100 μL of 5 × 10^5 CFU/mL bacteria in Mueller Hinton Broth (MHB, Hopebio, HB6231-1) or MHB (bacteria free) was added to a 96-well. 20 mg/mL peptide stock solution was added to the first well and given serial twofold dilutions to obtain final concentrations of 0.97, 1.95, 3.90, 7.81, 15.63, 31.25, 62.5, 125, 250, 500, and 1000 μg/mL. Plates were incubated at 37 °C for 20 h. Bacteria growth was monitored by measuring the OD_{600} using a microplate reader (Molecular Devices, Spectra Max M5). The MIC value was the minimum corresponding concentration at which the OD_{600} value was equal to the negative control (bacteria-free MHB). Then, 20 μL of the bacterial suspension or diluent was evenly spread on an LB plate (BD Diagnostics, 221994) and incubated at 37 °C overnight, followed by colony counting. After MIC assay, 20 μL of samples under greater than or equal to MIC concentrations were sub-cultured overnight at 37 °C on LB Agar plates. The MBC value was the minimum concentration corresponding to no bacteria growth.

Cell Viability

An MTT assay was employed to assess in vitro cytotoxicity of AMPs. NHDF and SHED cells were seeded a plated in 96 well plates at a density of 5000 cells/well and allow to adhere overnight at 37 °C, 5% CO2. The culture medium DMEM with 10% FBS and 1% P/S was replaced with culture medium containing 1, 2.5, 5, 10, 25, 50 μg/mL peptides. Blank medium was used as a negative control. After 24 h incubation period, cells were washed and added 10 μL of MTT solution and samples incubated for 3 h to form formazan. After that, the culture medium was replased with 100 μL DMSO for absolutely dissolved formazan. Absorbance was recorded at 570 nm using a microplate reader (Molecular Devices, Spectra Max M5). The percent viability was calculated as follows: (Absorbance peptide-treated cells-Absorbance blank) / (Absorbance untreated cells-Absorbance blank) ×100. GraphPad Prism 9 software was used to fit cytotoxicity curves and IC_{50} values were calculated with non-linear regression model.

Hemolytic Assay

Human whole red blood cells were centrifuged to obtain RBCs that were then washed and resuspend in PBS to 0.25% (v/v). In a 96-well plate, 75 μL of the RBCs solution was mixed with an equal volume of 2 × peptide dissolved in PBS to initiate the assay. Blank or 1% Triton - X100 containing PBS were used as negative and positive controls, respectively. Samples were incubated for 1 h with gentle shaking, followed by centrifugation at 2200 g for 10 min at 4 °C.100 μL of the supernatant from each well was added to an empty 96-well plate. The hemolysis was detected by absorbance at 415 nm using a microplate reader (Molecular Devices, Spectra Max M5), and calculated using the following equation:

$$((Absorbance_{treated\ hRBC} - Absorbance_{untreated\ hRBC}) \\ /(Absorbance_{triton-X100treated\ hRBC} - Absorbance_{untreated\ hRBC})) \times 100.$$

Circular Dichroism (CD) Spectroscopy

DOPC and DOPS were mixed equally and dried to a film with argon and subsequently lyophilized overnight. After that, suitable PBS was added and the liposome with a size of 100 nm (2.5 mM) was formed by ultrasound, vortex, and extrusion techniques. Peptide solutions (50 μM) were used to yield a peptide/lipid ratio of 1:50. CD wavelength spectra of the peptides were measured from 260 to 200 nm using a 1 mm path length quartz cell at 37 °C. Mean residue ellipticity, $[\theta]$, was calculated using the following equation: $[\theta] = (\theta obs/10lc)/r$, where θobs is the observed ellipticity in millidegrees, l is the length of the cell (cm), c is the concentration (M), and r is the number of residues.

Morphology study by SEM imaging

A suspension of bacteria in the exponential growth phase was incubated with AMPs (1 × MIC) for 12 h. A bacterial suspension without any treatment was used as the negative control. After centrifuging (1500 g, 12 min) and removing the supernatants, bacteria were fixed at 4 °C with PBS containing 2.5% glutaraldehyde (Sinopharm, 30092436). Samples were washed three times with PBS and then dehydrated using a series of ethanol solutions (30, 50, 70, 90, and 100%). The samples were sputter-coated with gold for observation using a Hitachi S-4800 field emission scanning electron microscope.

Effects of AMPs on bacterial membrane permeability

Bacteria were collected and washed three times with PBS. Then bacteria were resuspended to a working concentration ($OD_{600} = 0.1$) and treated with AMPs at 1 × MIC for 5 h at 37 °C. After staining with 20 μg/mL PI for 15 min, the samples were transferred to a glass slide for confocal imaging (FV1200, Olympus) or detected red fluorescence of PI on FL2-PE via flow cytometry (Accuri C6 Plus, Becton Dickinson). Seven pictures were taken for each treatment group, and the percentage of PI-stained bacteria was calculated using the following equation:

$$\text{PI} - \text{stained bacteria\%} = \text{Number of bacteria}_{\text{Stained with PI}} / \text{Number of bacteria}_{\text{Stained with DAPI}} \times 100. \quad (5)$$

Molecular dynamics simulations of peptide T1-2 with a model membrane

We used the original PDB ID: 1OMQ as a starting point for the peptide structure and mutated it to the current sequence (T1-2). The bacterial membrane was modeled using a composition of 60% POPC, 20% POPG, and 20% cholesterol, which mimics the key components of the membrane. At the beginning of the simulation, the peptide was placed onto the membrane with its helical axis parallel to the membrane plane. The simulations were performed for 500 ns using NAMD v2.14, with a 1.2 nm cut-off for interactions and 150 mM NaCl ion concentration. The CHARMM36m force field was used throughout the simulations[34,35]. The lipid order parameter (S_{zz}) was measure as $S_{zz} = <\frac{3\cos^2\theta - 1}{2}>$, where θ is the angle between the $C_{a-1} - C_{a+1}$ vector and the z-axis.

Membrane depolarization assay

The cytoplasmic membrane depolarization activity of the peptides was determined by measurements of fluorescence of the membrane-potential-sensitive dye $DiSC_3(5)$. Briefly, *E. coli* was centrifuged and washed three times with washing buffer (20 mM glucose, 5 mM HEPES, pH 7.2). Then resuspended to a working concentration ($OD_{600} = 0.05$) OD600 in the working buffer (20 mM glucose, 5 mM HEPES, 0.1 M KCl, pH 7.2). Thereafter, 100 μL bacteria were incubated with 20 nmol/L $DiSC_3(5)$ (53213-94-8, Aladdin) until a stable reduction of fluorescence was achieved (about 15 min). Samples were treated with 1 × MIC AMPs and membrane depolarization was then monitored by observing the change in the fluorescence emission intensity of $DiSC_3(5)$ ($\lambda_{ex} = 622$ nm, $\lambda_{em} = 670$ nm).

Permeabilization of outer membranes

The membrane permeability of the peptides was determined by using the NPN uptake assay. *E. coli* was grown to an OD600 of 0.5, centrifuged (9168 g at 4 °C for 10 min), washed, and re-suspended in buffer (5 mM HEPES, 5 mM glucose, pH 7.4) to a working concentration ($OD_{600} = 0.05$). Then, samples were treated with 10 μM NPN solution (P110559, Aladdin) in a 96-well plate. The background fluorescence was recorded at $\lambda_{ex} = 350$ nm and $\lambda_{em} = 420$ nm. 100 μl peptide solution was added to the 96-well plate and reached 1 × MIC as a final concentration. Fluorescence was recorded as a function of time until no further increase in fluorescence was observed.

Bacterial resistance development assays

For serial passage evolution, *S. aureus* was grown overnight in 3 mL MHB and diluted 1/10,000 by fresh MHB. 100 μL of bacteria were added to a 96-well plate, in the presence of varying concentrations of AMPs or ciprofloxacin (C861180, Maclin) at two-fold serial dilutions. Plates were incubated at 37 °C without shaking for 24 h, at which time they were read at 600 nm using a microplate reader (Molecular Devices, Spectra Max M5). After 24 h, bacteria that grew in the presence of the highest concentration of AMPs or ciprofloxacin were diluted 1/10,000 by fresh MHB, and once again introduced to varying concentrations of AMPs or ciprofloxacin at two-fold serial dilutions. This procedure was performed every 24 h over 25 days.

Antibiofilm assays

100 μL 10^6 CFU/mL bacteria suspension in MHB was treated with 1 × peptide solutions in a 96-well plate and incubated at 37 °C for 72 h. Then, the culturing medium in the wells was discarded and washed with PBS 3 times. Then fixed with 100 μl methanol (C861180, Sinopharm) for 30 min. The biofilm was air-dried and subsequently stained with 0.1% (w/v) crystal violet (C861180, Solarbio) for 30 min and then further washed with PBS. Thereafter the crystal violet was solubilized in 100 μL 30% acetic acid and the absorbance at 595 nm was measured with a microplate reader (Molecular Devices, Spectra Max M5).

Wound Mouse Model

Animal experiments were carried out complying with National Institutes of Health (NIH) guidelines for the Care and Use of Laboratory Animals, and the study protocol was approved by the Institutional Animal Care and Use Committee of Hunan University (HNU-IACUC-2021-102) under Dr. Junfeng Shi. Female ICR mice (6 to 8 weeks) were obtained from the SJA Laboratory Animal Co Ltd (Hunan, China). The mice were maintained under controlled conditions of temperature 22 °C ± 2 °C, relative humidity 50 ± 10%, and light–dark cycle 12 h.

Forty female ICR mice (6 to 8 weeks, each weighing 28 to 35 g) were anesthetized with an intraperitoneal injection of chloral hydrate (50 mg/kg). A 1.0 cm^2 open excision wound was created on the dorsal side of the skin of each mouse, reaching the depth of the loose subcutaneous tissue. *P. aeruginosa* suspensions (10^7 CFU) were inoculated on the excision wound area to establish the wound infection model. Mice were separated into five groups (PBS, T1-2, T1-5, T2-10, and levofloxacin), $N = 8$ mice per group. The treatment dose and frequency were 10 mg/kg and twice a day for 3 days, respectively. Treatments were started 1 h after infection by applying 60 μL of compound solutions or PBS to the wound area. On day 2, the bacterial load on the wound surface of the survived mice was determined by wiping the full wound surface with a sterile swab to transfer the bacteria to PBS and plating the PBS agar plate (supplemented with 20 μg/ml of ampicillin) for CFU determination. On day 3, mice were euthanized. The wound area and organ (liver, spleen and kidney) were collected for H&E staining analysis. Blood was withdrawn for evaluation of RBC parameters after treated with PBS, T1-2, T1-5, T2-10 and levofloxacin for 24 h by intraperitoneal injection. The statistical significance of differences in surface bacterial load was determined by the one-way ANOVA with Tukey's multiple comparison test...

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All data used in this paper are publicly available and can be accessed at https://github.com/jimmyrate/deepAMP. The source data for all figures and tables in the manuscript and in the Supplementary Information are provided in Source Data Files. Source data are provided with this paper.

Code availability

All codes and the trained models are available at https://github.com/jimmyrate/deepAMP.

References

1. Willyard, C. The drug-resistant bacteria that pose the greatest health threats. *Nature* **543**, 15–15 (2017).
2. Mancuso, G., Midiri, A., Gerace, E. & Biondo, C. Bacterial antibiotic resistance: the most critical pathogens. *Pathogens* **10**, 116 (2021).

3. Chivers, S. Why big pharma has abandoned antibiotics. *Nature* **586**, S50–S52 (2020).
4. Towse, A. et al. Time for a change in how new antibiotics are reimbursed: Development of an insurance framework for funding new antibiotics based on a policy of risk mitigation. *Health Policy* **121**, 1025–1030 (2017).
5. Hancock, R. E. & Sahl, H. G. Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. *Nat. Biotechnol.* **24**, 1551–1557 (2006).
6. Hancock, R. E. W., Alford, M. A. & Haney, E. F. Antibiofilm activity of host defence peptides: complexity provides opportunities. *Nat. Rev. Microbiol.* **19**, 786–797 (2021).
7. Fjell, C. D., Hiss, J. A., Hancock, R. E. W. & Schneider, G. Designing antimicrobial peptides: form follows function. *Nat. Rev. Drug Discov.* **11**, 37–51 (2012).
8. Derossi, D., Joliot, A. H., Chassaing, G. & Prochiantz, A. The third helix of the Antennapedia homeodomain translocates through biological membranes. *J. Biol. Chem.* **269**, 10444–10450 (1994).
9. Derossi, D. et al. Cell internalization of the third helix of the antennapedia homeodomain is receptor-independent*. *J. Biol. Chem.* **271**, 18188–18193 (1996).
10. Nielsen, E. J. B. et al. In vivo proof of concept of oral insulin delivery based on a co-administration strategy with the cell-penetrating peptide penetratin. *J. Control. Rel.* **189**, 19–24 (2014).
11. Jiang, K. et al. Discerning the composition of penetratin for safe penetration from cornea to retina. *Acta Biomater.* **63**, 123–134 (2017).
12. Ganz, T. et al. Defensins. Natural peptide antibiotics of human neutrophils. *J. Clin. Invest.* **76**, 1427–1435 (1985).
13. Zasloff, M. Magainins, a class of antimicrobial peptides from Xenopus skin: isolation, characterization of two active forms, and partial cDNA sequence of a precursor. *Proc. Natl Acad. Sci. USA* **84**, 5449–5453 (1987).
14. Zeng, X. et al. Deep generative molecular design reshapes drug discovery. *Cell Rep. Med.* **3**, 100794 (2022).
15. Huang, J. et al. Identification of potent antimicrobial peptides via a machine-learning pipeline that mines the entire space of peptide sequences. *Nat. Biomed. Eng.* **7**, 797–810 (2023).
16. Szymczak, P. et al. Discovering highly potent antimicrobial peptides with deep generative model HydrAMP. *Nat. Commun.* **14**, 1453 (2023).
17. Ma, Y. et al. Identification of antimicrobial peptides from the human gut microbiome using deep learning. *Nat. Biotechnol.* **40**, 921–931 (2022).
18. Das, P. et al. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nat. Biomed. Eng.* **5**, 613–623 (2021).
19. Witten, J. & Witten, Z. Deep learning regression model for antimicrobial peptide design. *bioRxiv*, https://doi.org/10.1101/692681 (2019).
20. Li, F. et al. Design of self-assembly dipeptide hydrogels and machine learning via their chemical features. *Proc. Natl Acad. Sci.* **116**, 11259–11264 (2019).
21. Yoshida, M. et al. Using evolutionary algorithms and machine learning to explore sequence space for the discovery of antimicrobial peptides. *Chem* **4**, 533–543 (2018).
22. Porto, W. F. et al. In silico optimization of a guava antimicrobial peptide enables combinatorial exploration for peptide design. *Nat. Commun.* **9**, 1490 (2018).
23. Das, P. et al. Pepcvae: Semi-supervised targeted design of antimicrobial peptide sequences. *arXiv preprint arXiv:1810.07743* (2018).
24. Becht, E. et al. Dimensionality reduction for visualizing single-cell data using UMAP. *Nat. Biotech.* **37**, 38–44 (2019).
25. Wyres, K. L., Lam, M. M. C. & Holt, K. E. Population genomics of Klebsiella pneumoniae. *Nat. Rev. Microbiol.* **18**, 344–359 (2020).
26. Osorio, D., Rondón-Villarreal, P. & Torres, R. Peptides: A package for data mining of antimicrobial peptides. *R. J.* **7**, 4 (2015).
27. Boman, H. G. Antibacterial peptides: basic facts and emerging concepts. *J. Intern. Med.* **254**, 197–215 (2003).
28. Eisenberg, D., Weiss, R. M. & Terwilliger, T. C. The helical hydrophobic moment: a measure of the amphiphilicity of a helix. *Nature* **299**, 371–374 (1982).
29. Dieltjens, L. et al. Inhibiting bacterial cooperation is an evolutionarily robust anti-biofilm strategy. *Nat. Commun.* **11**, 107 (2020).
30. Bai, S. et al. A polymeric approach toward resistance-resistant antimicrobial agent with dual-selective mechanisms of action. *Sci. Adv.* **7**, eabc9917 (2021).
31. Abramson, J. et al. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
32. Eisenberg, D., Weiss, R. M., Terwilliger, T. C. & Wilcox, W. Hydrophobic moments and protein structure. *Faraday Symp. Chem. Soc.* **17**, 109–120 (1982).
33. Pace, C. N. & Scholtz, J. M. A helix propensity scale based on experimental studies of peptides and proteins. *Biophys. J.* **75**, 422–427 (1998).
34. Phillips, J. C. et al. Scalable molecular dynamics with NAMD. *J. Comput. Chem.* **26**, 1781–1802 (2005).
35. Huang, J. et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73 (2017).

Acknowledgements

This project has been funded in whole or in part with federal funds from the National Cancer Institute, National Institutes of Health, under contract HHSN261201500003I to R.N. The content of this publication does not necessarily reflect the views or policies of the Department of Health and Human Services, nor does mention of trade names, commercial products, or organizations imply endorsement by the US Government. This research was supported [in part] by the Intramural Research Program of the NIH, National Cancer Institute, to R.N.

Author contributions

F.C. and X.Z. conceived the development of deepAMP and J.S. conceived the study of experimental validation. X.R. developed deepAMP and performed all computational experiments. T.L., J.S., D.Y., Z.W., Y.L., and J.S., performed all in vitro assays and in vivo mouse studies. Xiaoli L. and Xiaoyan L. performed partial computational experiments. Z.L. performed molecular dynamics simulation. R.N. and F.C. interpreted the data analysis. T.L. and X.R. drafted the manuscript. T.L., X.R., F.C., X.Z., R.N., and J.S. critically revised the manuscript. All authors critically revised and gave final approval of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-024-51933-2.

Correspondence and requests for materials should be addressed to Xiangxiang Zeng, Junfeng Shi or Feixiong Cheng.

Peer review information *Nature Communications* thanks Evandro Fang, and the other, anonymous, reviewers for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.