

DLFea4AMPGen de novo design of antimicrobial peptides by integrating features learned from deep learning models

Received: 11 November 2024

Accepted: 11 September 2025

Published online: 15 October 2025



Han Gao^{1,2,7}, Feifei Guan^{3,7}, Boyu Luo^{4,7}, Dongdong Zhang^{5,7}, Wei Liu^{6,7}, Yuying Shen³, Lingxi Fan³, Guoshun Xu¹, Yuan Wang¹, Tao Tu¹, Ningfeng Wu³, Bin Yao¹, Huiying Luo¹✉, Yue Teng⁴✉, Jian Tian¹✉ & Huoqing Huang¹✉

Deep learning models show promise in accelerating the design and optimization of antimicrobial peptides (AMPs), but current methods face challenges, such as low success rates, or large virtual library scales. In this study, we introduce DLFea4AMPGen, a bioactive peptide design strategy that leverages deep learning models to identify and extract key features associated with antimicrobial peptide activity. This approach enables the generation of peptide sequences with potential bioactivities. Using the SHapley Additive exPlanations (SHAP) method, we quantify the contribution of each amino acid in multifunctional peptides with potential antibacterial, antifungal, and antioxidant activities. Key feature fragments (KFFs) with the highest average contributions are extracted and classified into four subfamilies based on amino acid frequency. These high-frequency amino acids are systematically arranged to generate a plausible sequence subspace for candidate peptides, from which 16 representative sequences were selected for experimental validation. The results show that 75% (12/16) of the sequences exhibited at least two types of activity. Notably, D1 exhibits broad-spectrum antimicrobial activity, including efficacy against multidrug-resistant clinical pathogenic isolates both in vitro and in vivo. This proof-of-concept study underscores the potential of the DLFea4AMPGen platform for efficient design and screening of bioactive peptides, showcasing its value in AMP research.

AMPs have emerged as promising candidates for combating antibiotic-resistant bacteria and have consequently received increasing research attention^{1–5}. Their unique mode of action, typically involving the physical disruption of bacterial membranes, offers a potential means for circumventing existing mechanisms of resistance^{6–8}. Historically, AMP

design has primarily focused on the discovery of existing natural peptides^{9–12} or the modification of natural peptides to enhance their bioactivity^{13,14}.

Artificial intelligence is currently revolutionizing the field of peptide drug design, especially through advances in large model

¹State Key Laboratory of Animal Nutrition and Feeding, Institute of Animal Science, Chinese Academy of Agricultural Sciences, Beijing, China. ²School of Life Sciences, Inner Mongolia University, Hohhot, Inner Mongolia, China. ³National Key Laboratory of Agricultural Microbiology, Biotechnology Research Institute, Chinese Academy of Agricultural Sciences, Beijing, China. ⁴State Key Laboratory of Pathogen and Biosecurity, Academy of Military Medical Sciences, Beijing, China. ⁵Western Medical Branch of PLA General Hospital, Beijing, China. ⁶Laboratory Department of the Second Medical Center of the General Hospital of the PLA, Beijing, China. ⁷These authors contributed equally: Han Gao, Feifei Guan, Boyu Luo, Dongdong Zhang, Wei Liu.

✉ e-mail: luohuiying@caas.cn; yueteng@me.com; tianjian@caas.cn; huoqinghuang@126.com

technologies^{15–17}. By leveraging deep learning models, it is now possible to rapidly design a multitude of AMPs that have not been found or may not exist in nature¹⁸. Currently, de novo AMP design primarily follows two approaches: First, an increasing number of generative deep learning models have been developed to generate protein and peptide sequences. Generative models, such as variational autoencoders^{19,20}, Chroma²¹, and ProGen²², are trained on unlabeled data to suggest new protein sequences based on an understanding of the natural protein sequence landscape. Second, considerable progress has been made in utilizing deep learning models to explore AMPs from the entire virtual library of peptides. By scanning a virtual library, AMP prediction models can accurately identify sequences with the strongest potential for antimicrobial activity^{23,24}. Both approaches have successfully generated potent AMPs for targeting drug-resistant pathogens, thus broadening the spectrum of available bioactive peptides and accelerating the discovery of functional peptides beyond those in natural reservoirs.

Despite the considerable success of these current de novo design methods, they are still accompanied by some obvious shortcomings that limit their efficiency and require further improvement. In particular, there is a need for highly effective strategies to screen high-quality active peptides from the large virtual libraries generated by these generative deep learning models in order to improve the success rate of validation^{19,20}. By developing and combining different deep learning models to explore the entire virtual library, the success rate of de novo design has been significantly improved. However, existing traversal design strategies require listing all possible peptide sequence combinations, resulting in a large number of generated sequences²³. For example, the plausible sequence subspace of a 13-amino-acid (13-AA) peptide fragment contains 20¹³ sequences, which incurs a substantial computational burden. Database filtering techniques can be used to obtain a limited set of AAs, significantly reducing the scale of the entire virtual library^{24,25}, but result in AMP low sequence diversity. Additionally, bacterial infections can induce oxidative stress in the host, leading to host cell damage and worsening tissue damage^{26–28}, a factor that is often overlooked in early peptide drug screening. Therefore, AMP design should also consider activities that mitigate inflammation arising from oxidative stress, which has been demonstrated in previous studies^{29–31}.

In response to the challenges outlined above, we present DLFea4AMPGen for the generation of de novo AMPs with high rates of bioactivity. Peptides with potential antibacterial, antifungal, and antioxidant properties were identified by each respective model. KFFs that played the most important roles in these peptides were extracted through the SHAP method and categorized into four subfamilies based on phylogenetic analysis, each displaying unique AA frequency patterns. By rearranging the most common AAs at each position in each subfamily, we generated four comprehensive sequence spaces from which 16 representative candidate AMPs (c_AMPs) were selected for experimental validation. Our results showed a 75% positive rate for antimicrobial activity, with several peptides exhibiting all three bioactivities. Notably, D1 and D2 demonstrated broad-spectrum antibacterial activity against various bacterial strains, including drug-resistant ones. Furthermore, in vivo studies revealed that D1 treatment could reduce bacterial load and alleviate inflammatory response in sepsis model mice. This study showcases how deep learning-based bioactivity prediction can facilitate the precise design of peptides with enhanced bioactivity.

Results

Design of a feature-based strategy for multifunctional AMP generation

To establish a deep learning model framework for predicting the potential antibacterial, antifungal, and antioxidant activities of peptides, we used existing bioactive peptide datasets published in a

previous study³² to fine-tune the pre-trained Mindspore proteinBERT (MP-BERT) model³³ (Fig. 1a and Supplementary Fig. 1), resulting in three bioactive peptide models, which were antibacterial peptides for MP-BERT, antifungal peptides for MP-BERT, and antioxidant peptides for MP-BERT (ABP-MPB, AFP-MPB, and AOP-MPB). What's more, a total of 20 bioactive peptide datasets with 18 distinct activities (Supplementary Fig. 2 and Supplementary Table 1) in that previous study were considered as a candidate peptide set with potential triple activities.

In this strategy, the multi-stage pipeline was used to generate peptides with antibacterial, antifungal, and antioxidant activities (Fig. 1b). We first employed the ABP-MPB, AFP-MPB, and AOP-MPB models to predict peptides in all 20 of the above-mentioned bioactive peptide datasets, then selected only those peptides predicted to be positive for bioactivity by all three models. Here, the SHAP method is used to extract the features identified by these deep learning models, with each amino acid assigned a different SHAP value. To eliminate possible interference from background AAs, for each peptide sequence, we only extracted one 13-AA fragment that had the highest sum of average SHAP values for all three models, based on the principle that these so-called KFFs had the greatest impact on peptide bioactivity. We then constructed a phylogenetic tree to further classify homologous KFFs into four subfamilies based on distinct amino acid features. For each subfamily, a plausible sequence subspace was then constructed by identifying the most frequently occurring residues at each position, and then systematically organizing these high-frequency AAs into every possible sequence combination. Finally, we selected representative sequences from each plausible sequence subspace for in vitro and in vivo experimental validation (Fig. 1c, d).

Developing deep learning models for bioactive peptide identification

To construct an accurate prediction model for bioactive peptides, a transfer learning approach was utilized to fine-tune the pre-trained MP-BERT protein model, which was developed in our previous work³³. The architecture generally comprised multi-head attention mechanisms, as well as normalization and feed-forward networks implemented in a repetitive manner across six encoding layers (Supplementary Fig. 1). The progressive six hidden layers in each model showed more and more pronounced differentiation between positive and negative samples from layer 1 to layer 6 (Supplementary Fig. 3), indicating the model's capacity to extract informative features for sample classification. As displayed in Fig. 2a, b, on a held-out testing set, it can be concluded that ABP-MPB, AFP-MPB, and AOP-MPB models achieve the most superior results. Moreover, even when evaluated on entirely new datasets not used during fine-tuning, the model maintains a prediction accuracy of over 84%, demonstrating strong generalization capability (Supplementary Fig. 4). In terms of evaluation metrics including accuracy, precision, recall, F1 score, and area under the curve (AUC), the performance of our model consistently surpasses that of traditional machine learning models such as Support Vector Machine (SVM) and eXtreme Gradient Boosting (XGBoost), as well as deep learning models like Convolutional Neural Network (CNN) and two other state-of-the-art (SOTA) approaches for AMP prediction^{32,34}. Furthermore, across all 20 bioactive peptide datasets, the model fine-tuned based on MP-BERT outperformed UniDL4BioPep (Supplementary Fig. 5), which had previously shown superior performance over other existing bioactive peptide prediction models.

To identify multifunctional peptides with antibacterial, antifungal, and antioxidant properties, we utilized ABP-MPB, AFP-MPB, and AOP-MPB models to predict bioactivity in a total of 23,346 peptides from 20 different bioactive peptide datasets. Among them, 4760 were predicted to possess all three bioactivities (Fig. 2c). As expected, most of these 4760 “triple activities” peptides were sourced from datasets originally linked to some form of antimicrobial activity, such as ABP,

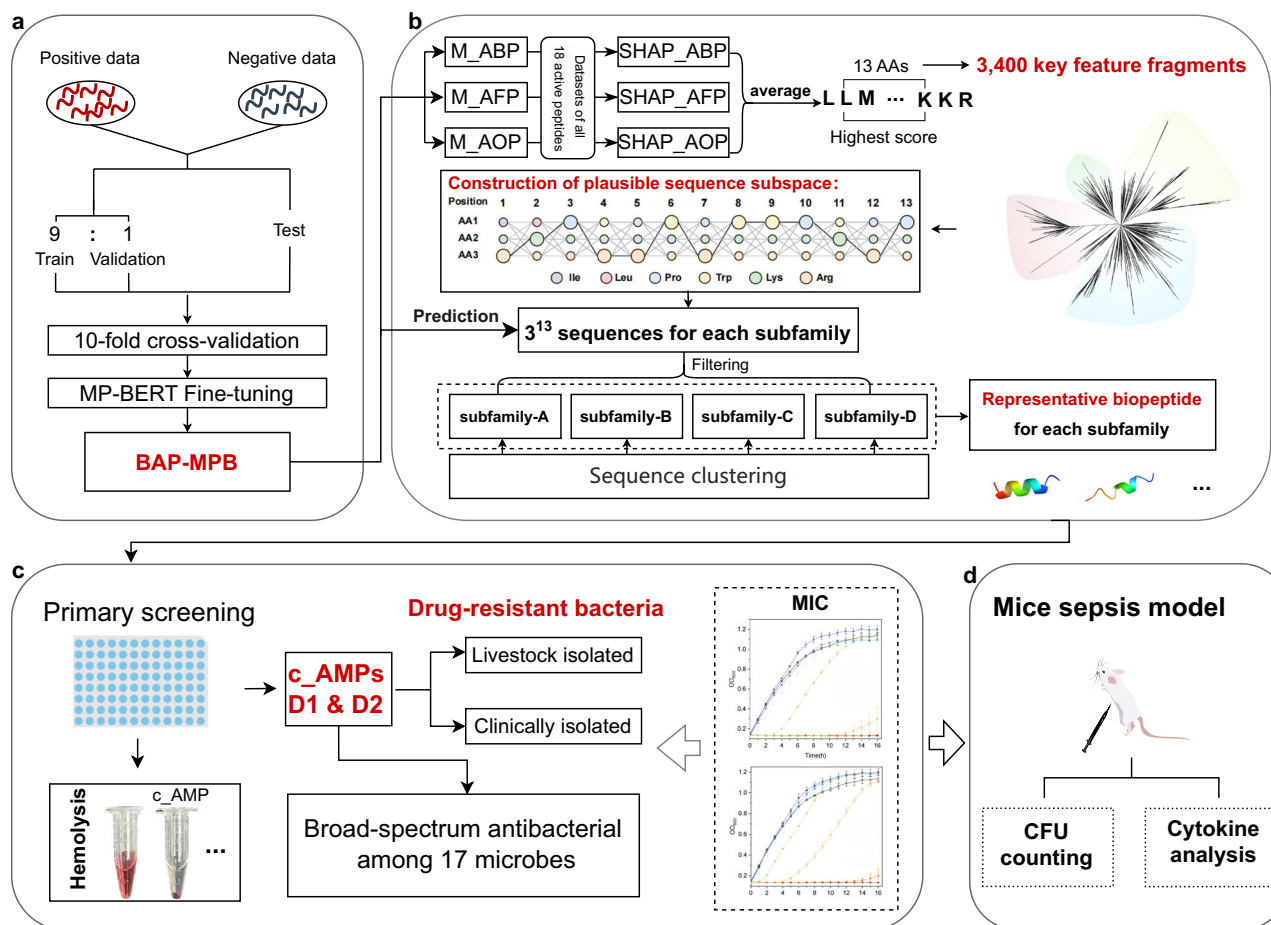


Fig. 1 | Schematic of the study workflow. **a** We began with model construction, using bioactive peptide datasets from a previous study. By fine-tuning the pre-trained MP-BERT model, we mainly developed three models, which were ABP-MPB, AFP-MPB, and AOP-MPB, collectively referred to as BAP-MPB (Bioactive peptide for MP-BERT). Furthermore, these three models were used to predict peptides with potential triple activities. **b** Subsequently, based on the SHAP interpretation of these three models, we applied a 13-AA sliding window to identify KFFs with the highest average SHAP value for each peptide sequence that were predicted to be positive for bioactivity by all three models. Next, with distinct AA features, these

KFFs were divided into four subfamilies, and the top three high-frequency AAs at each position were systematically organized into every possible sequence combination to form plausible sequence subspaces. Representative sequences from each plausible sequence subspace were selected as c_AMPs for chemical synthesis. **c, d** Finally, we conducted further efficacy tests and mechanistic analyses for antimicrobial and antioxidant activities in vitro ($n = 3$ biologically independent replicates in the MIC assays, mean $\pm$ s.d.), as well as in vivo experiments using a mouse sepsis model.

AMP, ACP, and so on (Fig. 2d). Subsequently, we focused on these 4760 sequences with potential triple activities for further analysis.

Extraction of important AAs and key feature fragments based on the SHAP method

To improve the interpretability of each model and potentially gain insights into the molecular mechanisms of each predicted bioactivity, we employed the SHAP method³⁵ to quantify the contribution of each AA at each position. Every AA is treated as an independent feature and assigned a distinct SHAP value. AA features differ across peptides with various activities, while antimicrobial-related AA features are similar (Supplementary Fig. 6). According to statistics, Cys, Trp, Lys, Pro and Arg in ABP; Cys, Arg, Lys, Trp, and His in AFP; and His, Trp, Tyr, Cys, and Pro in AOP were the most likely to make positive contributions to each respective bioactivity (Supplementary Fig. 7). To integrate these features into one sequence, we normalized and then averaged SHAP values from the ABP-MPB, AFP-MPB and AOP-MPB models for every AA. As shown in Fig. 3a, the 13-AA length is the most common length that is predicted to exhibit all three activities. Therefore, for each peptide sequence, we only extracted one fragment with the highest cumulative average SHAP value using a 13-AA sliding window to eliminate possible interference from background AAs, and designated it as

KFF (Fig. 3b). Excluding peptides shorter than 13 AAs or KFF with a negative sum, a total of 3400 KFFs were extracted. By analysis, the proportion of Lys, Cys, Gly, Trp, Arg, Pro, Ala, His, and Leu was all higher in the KFFs dataset than that in 20 bioactive peptide datasets, with an improvement rate ranging from 0.14% to 4.11% (Fig. 3c), which suggested that KFFs integrate and highlight AA features associated with three bioactivities.

These 3400 KFFs were then clustered into a phylogenetic tree based on sequence similarity, which revealed four distinct subfamilies, each with unique AA distributions (Fig. 3d). After aligning all KFFs in each subfamily, the top three AAs at each position, which together accounted for 31.31% to 54.38%, were considered fundamental elements for each KFF subfamily and were assembled into different combinations of the plausible sequence subspace (i.e., 3¹³ combinations per subfamily). As shown in Fig. 3e, the primary AAs in subfamily-A were Gly, Ala, Leu, Cys, and Lys (with Ala and Cys each appearing at only one position); Gly, Ala, Leu, and Lys in subfamily-B; Gly, Leu, Cys, and Lys in subfamily-C (with Leu appearing at two positions); and Leu, Ile, Trp, Pro, Lys, and Arg in subfamily-D (with Leu and Ile each appearing at only one position). Leveraging SHAP values to define core AA features thus yielded high-confidence sequence spaces of 3¹³ sequences for each subfamily.

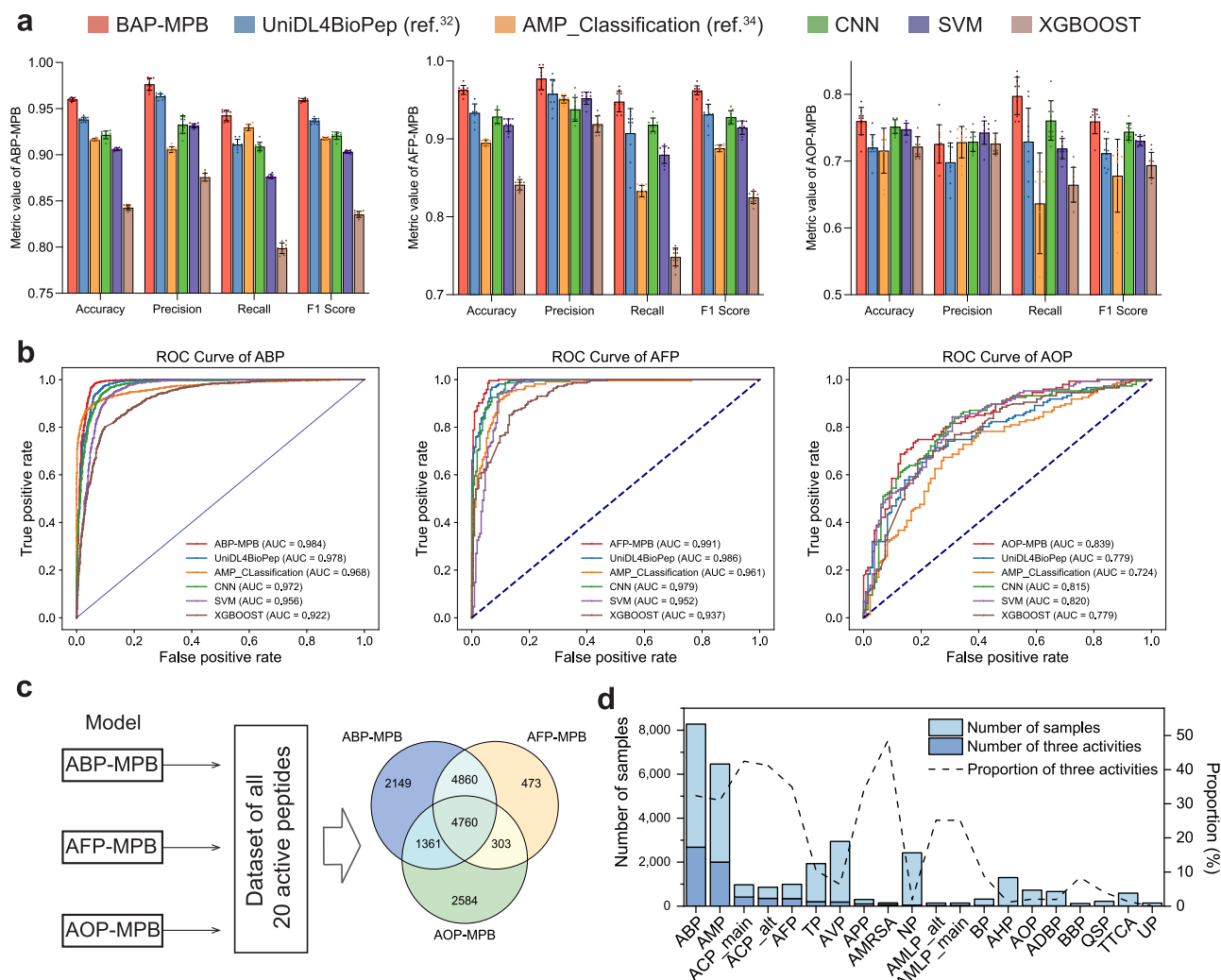


Fig. 2 | Comparative performance of different models and activity prediction with multifunctionality analysis. **a** The predicted performance on the test set, based on results from 10-fold cross-validation (mean $\pm$ s.d. over 10 folds, $n = 10$). Accuracy, precision, recall, and F1 scores were obtained by the trained BAP-MPB (which are ABP-MPB, AFP-MPB, and AOP-MPB) and the other models^{32,34}, all of which were trained and evaluated on the same dataset. For all four metrics, higher scores indicate better performance. **b** ROC curves of six models on the ABP, AFP, and AOP datasets, and the area under the curve is AUC. **c** Predictions made by the

ABP-MPB, AFP-MPB, and AOP-MPB models on all 20 datasets, retaining only positively predicted samples. Finally, the intersection of positive predictions from the three models is shown, with the number 4760 at the center representing peptides predicted to be bioactive by all three models. **d** The number of positive samples in each original bioactive peptide dataset, along with the number of samples that predicted with antibacterial, antifungal, and antioxidant activities (bar chart, left Y-axis), and the proportion of samples with triple activities in each dataset (line chart, right Y-axis).

Formation of plausible sequence subspaces and identification of representative AMPs

After generating a plausible sequence subspace of 3^{13} , or 1,594,323, sequences for each KFF subfamily, we screened c_AMPs by filtering out relatively low-quality sequences, retaining only those with probability >0.99 in all three models. Based on previous studies³⁶, sequences with less than two positive charges were excluded. Consequently, these plausible sequence subspaces were reduced to 893,093, 273,615, 127,427, and 666,165 sequences for the subfamily-A, subfamily-B, subfamily-C, and subfamily-D, respectively (Fig. 4a).

Given the limitations of traditional motif extraction methods for handling short sequences (Supplementary Fig. 8), we employed an improved strategy. In order to maximize diversity among the remaining sequences, sequences with high similarity were grouped together within each plausible sequence subspace, and the representative sequences were selected from each group for further analysis. By applying the Elbow method in k-means clustering, each plausible sequence subspace was divided into four clusters (Supplementary Fig. 9), with the sequence nearest to the cluster center

designated as the representative sequence. As a result, c_AMPs from subfamily-A (Fig. 4b) and subfamily-B (Fig. 4c) were predominantly composed of α -helix, random coil peptides, and the mixture of these two, while those from subfamily-C (Fig. 4d) primarily contained a β -sheet and random coil peptides, and peptides in subfamily-D (Fig. 4e) were generally random coil peptides. We generated 16 sequences through de novo design, none of which are found in existing databases (Supplementary Fig. 10). These sequences were further evaluated through experimental validation.

Preliminary evaluation of 16 c_AMPs in vitro

To investigate their possible antibacterial, antifungal, and antioxidant activities, these 16 c_AMPs were synthesized and validated (Fig. 5a and Supplementary Table 2). Antibacterial assay was conducted at a peptide concentration of 128 μ M, testing three Gram-negative bacterial strains (*Escherichia coli* O157:H7, *Klebsiella pneumoniae* ATCC700603, *Salmonella typhimurium* ATCC14028) and three Gram-positive bacterial strains (*Micrococcus luteus* ATCC4698, *Bacillus subtilis* WB600, *Staphylococcus aureus* ATCC6538). Among

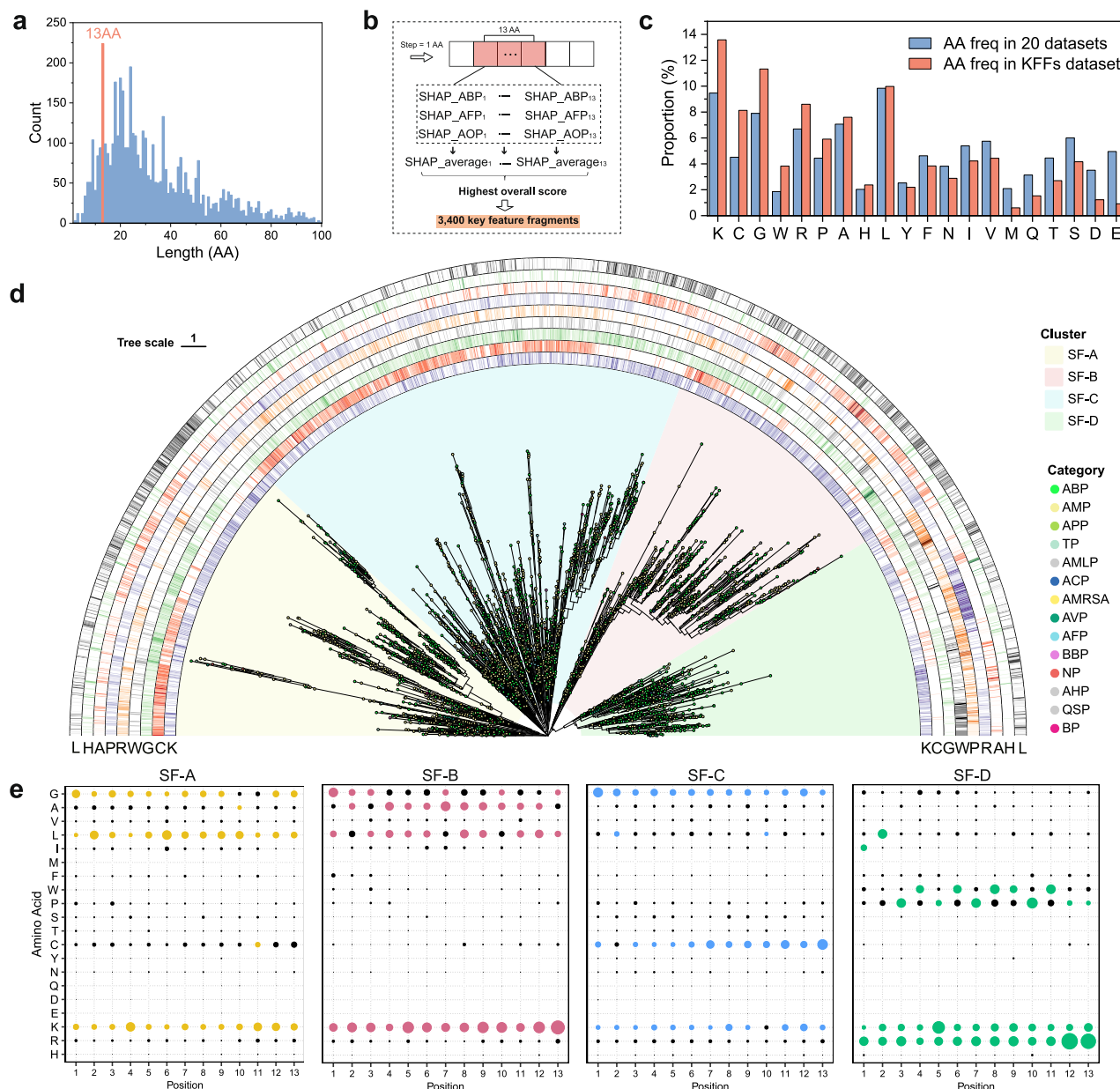


Fig. 3 | Key AA features and key feature fragments. **a** The length distribution of 4760 samples with potential antibacterial, antifungal, and antioxidant activities, with sequences of 13 AAs appearing most frequently. **b** Illustration of KFFs extraction. A window of 13 AAs is set to slide across the entire sequence with a step size of 1 AA. Fragments with the highest sum of average SHAP values and scores greater than 0 are selected. 3400 fragments that meet all conditions are being focused on as KFFs. **c** Comparison of the proportion of each AA in KFFs and in all

positive samples across 20 bioactive peptide datasets. **d** Distribution of 3400 KFFs in four subfamilies observed across the phylogenetic tree, which were subfamily-A (SF-A), subfamily-B (SF-B), subfamily-C (SF-C), and subfamily-D (SF-D). The colored dots at the tips of the branches indicate the original bioactive peptides source of each KFF. **e** AA distribution at the 13 positions in each KFF subfamily, and the highlighted AAs at each position are the top three AAs that most frequently occur.

Gram-negative bacteria, five treatment groups exhibited significant inhibition, which were B4, D1, D2, D3, and D4, with D1, D2, and D3 showing the highest inhibition rates (range: 75%–100%). For Gram-positive bacteria, 10 c_AMPs showed inhibitory effects towards at least one bacterial strain, and nine of which resulted in 100% inhibition. More specifically, A1, A3, B2, C1, C2, and D1–D4 could completely inhibit *M. luteus* ATCC4698 growth; C1, C2, and D1–D4 negatively impacted *B. subtilis* WB600; and D1 and D2 showed inhibitory effects in *S. aureus* ATCC6538 cultures. In the treatment of the fungal strain *Candida albicans* ATCC10231 with 128 μ M peptides, B2, C2, and D1–D4 all achieved complete inhibition of growth, reaching 100% inhibition. Additionally, ABTS⁺ free radical scavenging assays

with 1 mg/mL c_AMPs showed that A3 and C1–C4 had $\geq 90\%$ free radical scavenging rates, while D1, D2, D3, and D4 had scavenging rates of 72%, 54%, 62%, and 72%, respectively. These results suggested that c_AMPs D1 and D2 had the strongest broad-spectrum effects and exhibited all three bioactivities.

Furthermore, the growth curves demonstrated that certain peptides exhibited no inhibitory effects on microorganism growth at 16 h, but were able to inhibit bacterial or fungal growth during the earlier logarithmic growth phases (Supplementary Fig. 11). For example, A3 and B2 effectively halted the growth of *B. subtilis* WB600 for up to 3 h and 6 h post treatment, while C4 demonstrated the ability to inhibit the growth of *C. albicans* ATCC10231 for up to 2 h post treatment.

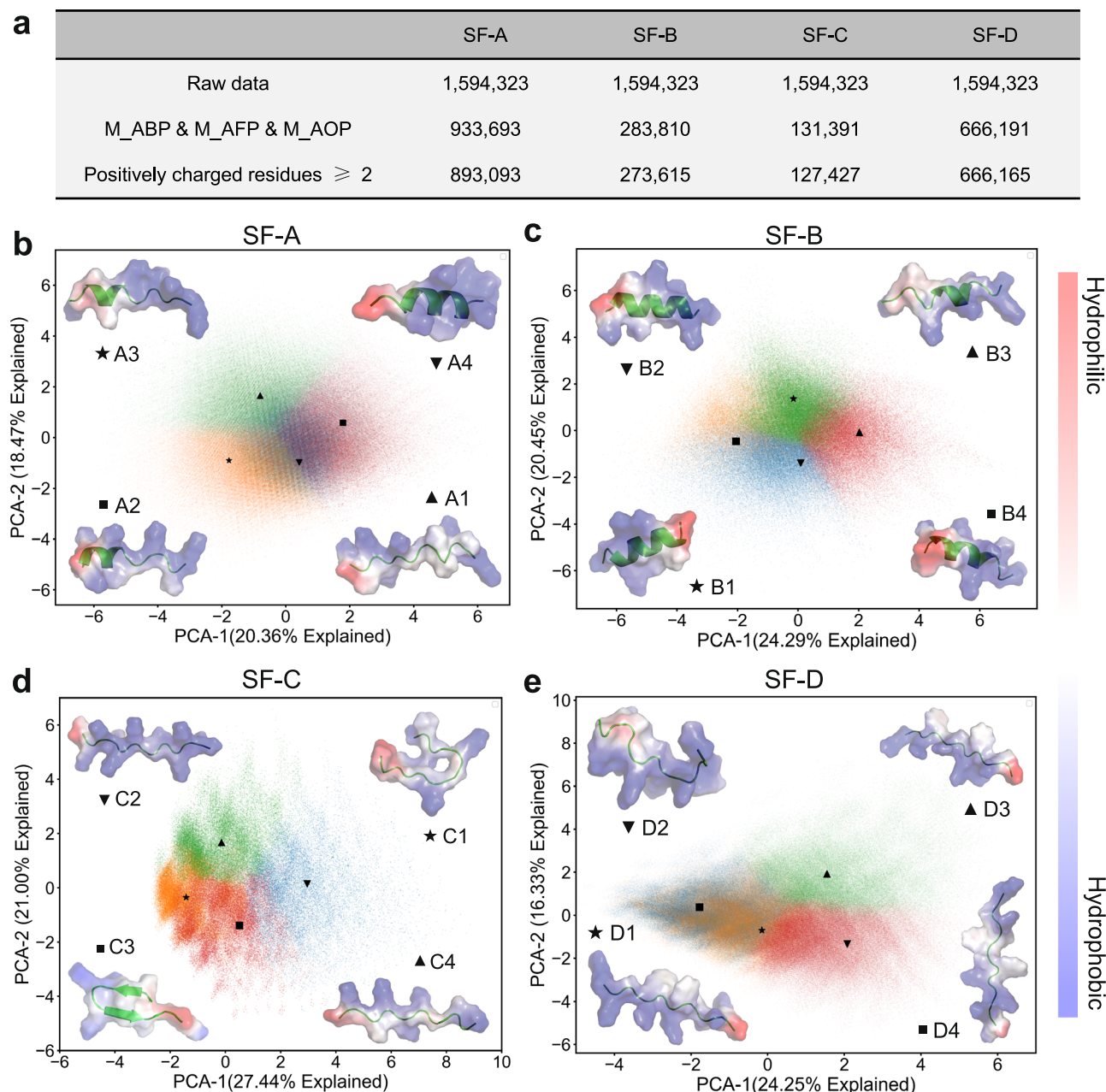


Fig. 4 | Screening in plausible sequence subspaces. **a** The number of remaining sequences in subfamily-A (SF-A), subfamily-B (SF-B), subfamily-C (SF-C), and subfamily-D (SF-D) after one-by-one step of stringent selection criteria, respectively. **b–e** Clustering within each subfamily and the selection of representative sequences. In the PCA plot, different colored dots represent different clusters, with

the black-labeled dots indicating the position of the selected representative sequence. The surrounding structural diagrams depict the three-dimensional structure of each selected representative sequence: the inner part shows the detailed three-dimensional structure, while the outer semi-transparent part illustrates the peptide's hydrophobicity distribution.

Additionally, we assessed the safety profile of all these 16 candidate peptides at a concentration of 128 μ M by evaluating their hemolytic activity against erythrocytes from rabbits and cytotoxicity on 3T3 cells. Notably, 15 peptides had hemolysis rates of less than 4%, indicating their potential for safe clinical application. Only B2 exhibited a hemolysis rate of 6.57%, and peptide B3 showed a cytotoxicity rate of 6.97% (Fig. 5b).

Further evaluation of D1 and D2 against drug-resistant strains
Given their broad-spectrum antimicrobial activity in vitro, the inhibitory effects of D1 and D2 were subsequently evaluated against antibiotic-resistant strains isolated from livestock or clinical environments. D1 exhibited a minimum inhibitory concentration (MIC) of 4 μ M

against the antibiotic-resistant livestock strain *E. coli* z44, while D2 had MIC values of 8 μ M (Fig. 5c). Impressively, at a concentration of 2 μ M, D1 was able to inhibit bacterial growth for up to 4 h compared to untreated control cultures. In the case of the clinical antibiotic-resistant strain, *S. aureus* 09057, D1 showed an MIC of 16 μ M and was able to inhibit the growth of both strains for 11 h at 8 μ M, while D2 had an MIC of 32 μ M (Fig. 5d). Additionally, further testing on other drug-resistant bacteria strains (Supplementary Figs. 12 and 13) confirmed the broad-spectrum antibacterial activity of D1 and D2. Among them, D1 demonstrates a slightly superior inhibitory effect compared to D2. Notably, after 100 generations of continuous treatment with D1, no resistance could be detected in *S. aureus* 09057 (Supplementary Fig. 14).

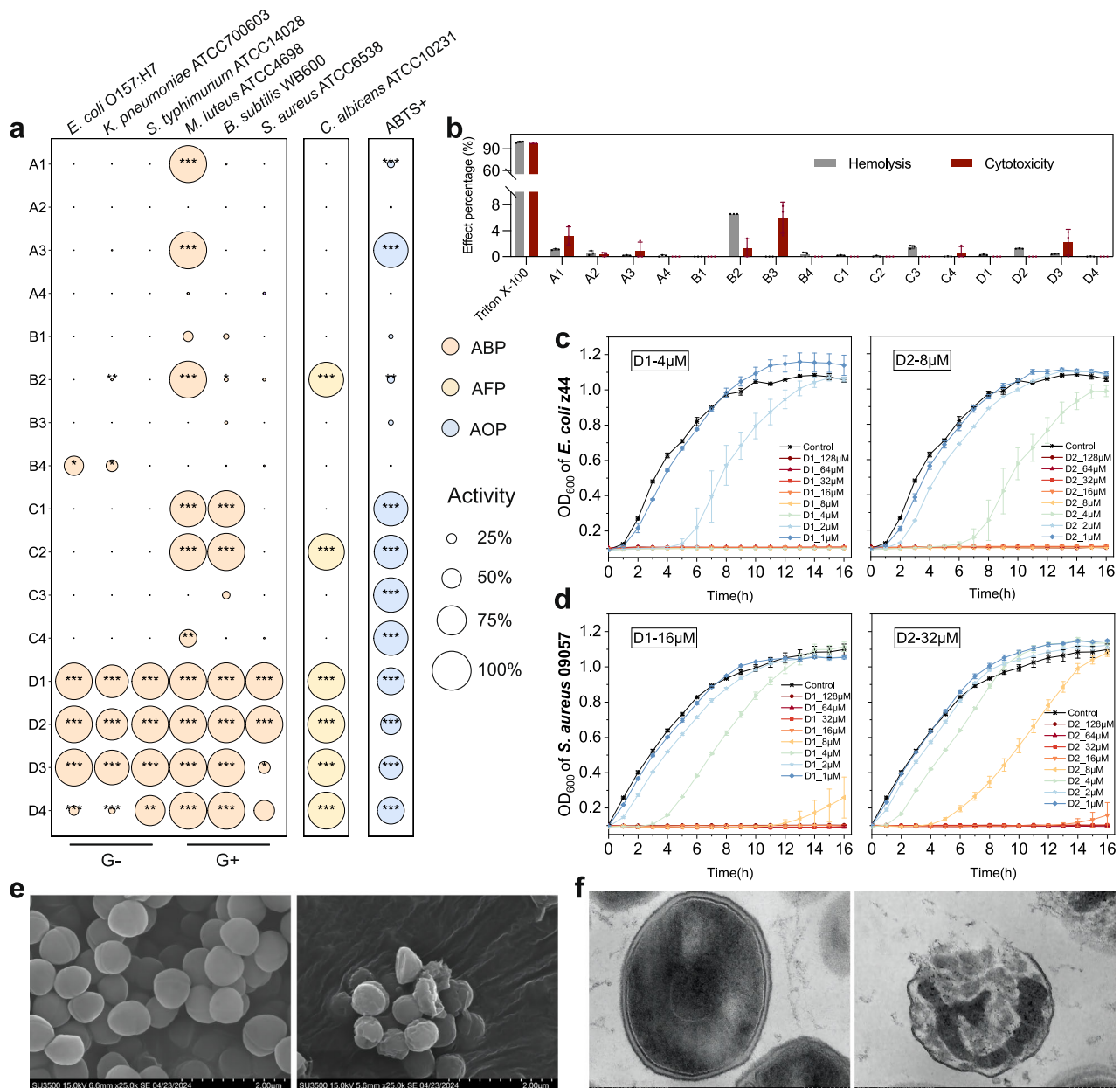


Fig. 5 | Experimental validation and potency assays of cAMPs. a Inhibition rates of six bacterial and one fungal strain at 128 μM , and the scavenging rate of ABTS⁺ radicals at 1 mg/mL. Larger circles indicate better activity. Statistical significance was assessed using two-sided Dunnett's test with correction for multiple comparisons: '*' means $0.01 < p < 0.05$; '**' means $0.001 < p \leq 0.01$; and '***' means $p \leq 0.001$. **b** The hemolytic and cytotoxic activities of 16 cAMPs were evaluated at a concentration of 128 μM . Each group included three biologically independent replicates ($n = 3$), with data presented as mean $\pm$ s.d. Triton X-100 served as a positive control, while PBS was used as a negative control for hemolytic activity, and DMEM was the control for cytotoxicity. The growth curves of representative drug-resistant bacteria, which were *E. coli* z44 (**c**) isolated from sick chickens in livestock and *S.*

aureus 09057 (**d**) from clinical environments. Each group included three biologically independent replicates ($n = 3$), with data presented as mean $\pm$ s.d. The control group demonstrated normal bacterial growth. The concentration gradient of D1 and D2 was set from 128 μM to 1 μM in two-fold dilutions, with different colors representing each concentration. The concentration indicated in each panel is the MIC of the peptide against each bacterial strain. **e** SEM images of *S. aureus* 09057 treated with PBS and D1 (64 μM) at low magnification (scale bar, 2 μm). **f** TEM images of *S. aureus* 09057 treated with PBS and D1 (64 μM) at low magnification (scale bar, 200 nm). Images shown are representative of at least three independent experiments with similar results. Additional images and magnifications are provided in Supplementary Fig. 15, confirming the reproducibility of the observations.

To further investigate whether D1 affected the cellular integrity, we conducted scanning and transmission electron microscopy (SEM and TEM) in *S. aureus* 09057 cells treated with 64 μM D1. SEM showed that, as processing time increases, obviously ruptured cells with compromised membrane integrity relative to the intact cells in the untreated control group (Fig. 5e). Similarly, TEM also clearly depicted damage to the cellular membranes and loss of cytoplasmic contents in D1-treated cells but not the untreated controls (Fig. 5f and Supplementary Fig. 15).

Therapeutic efficacy in treating bacterial infection in vivo

Given the above in vitro results, we next examined whether D1 could confer therapeutic effects in a mouse sepsis model in vivo (Fig. 6a). Mice were infected with the high-risk, antibiotic-resistant clinical isolate *S. aureus* 09057. At 1 h post-infection, the infection model was confirmed to be successfully established (Supplementary Fig. 16). Mice were then systemically treated with 200 μL of D1 (20 mg/kg, intraperitoneal injection). At 12 h post-infection, the mice were sacrificed,

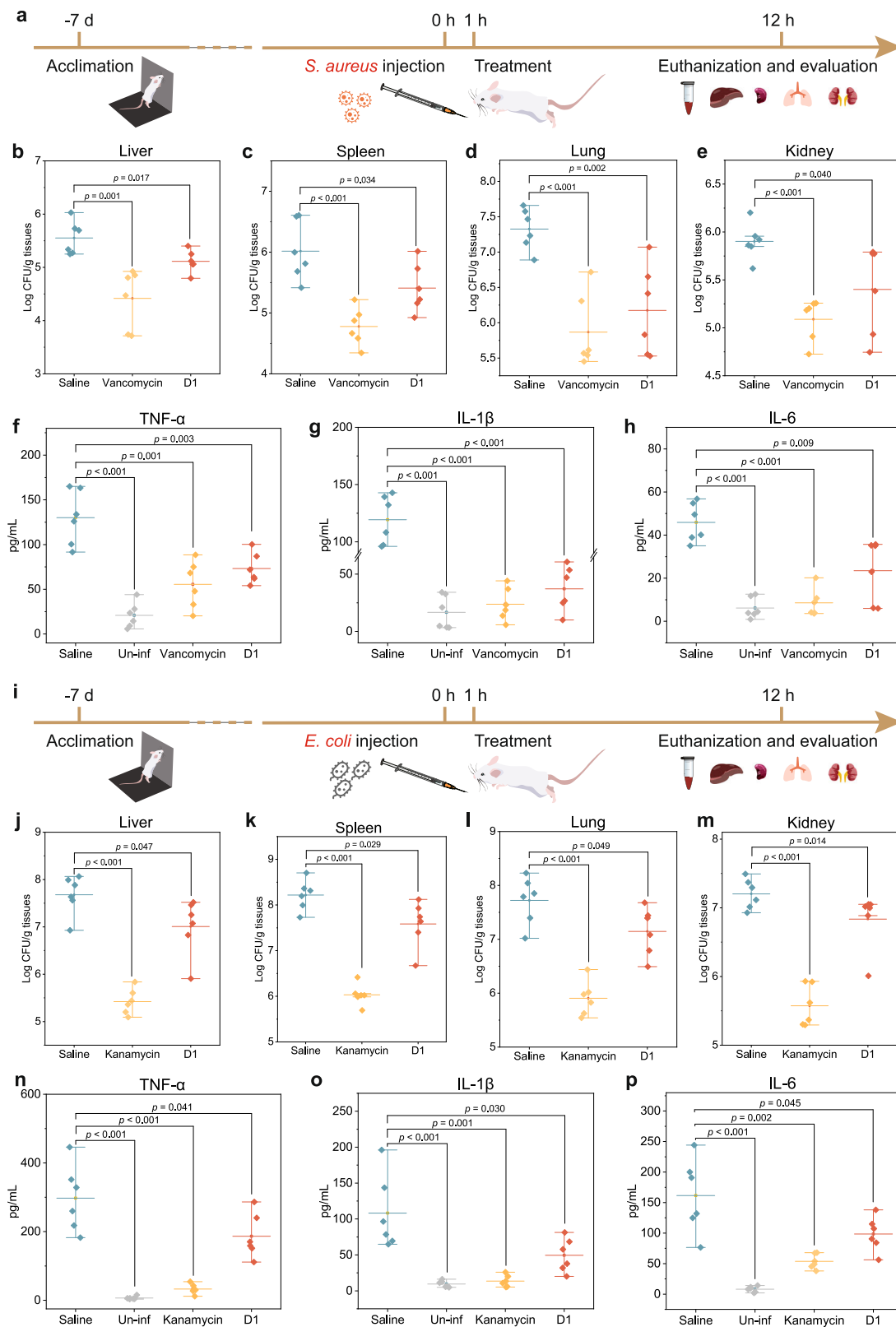


Fig. 6 | Therapeutic efficacy in vivo. **a** Schematic diagram of the experimental setup for the mice systemic infection model using *S. aureus* 09057. **b–e** Bacterial loads of *S. aureus* 09057 in liver, spleen, lung, and kidney of uninfected mice, and of infected mice after treatment with saline, vancomycin, and D1. **f–h** Serum levels of TNF- α , IL-1 β , and IL-6 in healthy mice and *S. aureus*-infected mice treated with saline, vancomycin, and D1. **i** Schematic diagram of the experimental setup for the

mice systemic infection model using *E. coli* z44. **j–m** Bacterial loads of *E. coli* z44 in liver, spleen, lung, and kidney of uninfected mice, and of infected mice after treatment with saline, kanamycin, and D1. **n–p** Serum levels of TNF- α , IL-1 β , and IL-6 in healthy mice and *E. coli*-infected mice treated with saline, kanamycin, and D1. Statistical analysis was conducted using a two-sided one-way ANOVA. Data represent mean $\pm$ s.d. of $n = 6$ biologically independent replicates.

autopsied, and organs were collected for quantification of bacterial load. Colony counts showed that bacterial loads were significantly lower in the liver, kidney, spleen, and lungs of vancomycin-treated mice compared to the corresponding organs of untreated sepsis control mice, indicating the successful establishment of the infection model. Compared to the saline treatment, bacterial loads in the liver, kidney, spleen, and lungs were significantly reduced by an average of 0.44–1.15 log in CFU/g after treatment with D1 (Fig. 6b–e). Moreover, compared with the saline treatment group, the levels of pro-inflammatory factors tumor necrosis factor- α (TNF- α), interleukin-1 β (IL-1 β), and interleukin-6 (IL-6) in the serum of mice in the D1-treatment group were significantly reduced (Fig. 6f–h), suggesting that D1 could also alleviate inflammatory response in vivo.

In the sepsis model induced by *E. coli* z44 infection (Fig. 6i), under the same treatment conditions, D1 treatment significantly reduced the bacterial load in the visceral organs of mice by an average of 0.47–0.67 log CFU/g (Fig. 6j–m). Moreover, compared to the untreated group, D1 significantly decreased the levels of pro-inflammatory cytokines in the mice (Fig. 6n–p), indicating that D1 can alleviate the inflammatory response in mice with *E. coli* z44-induced sepsis.

Discussion

Deep learning models can be powerful tools for predicting protein function, but the learned relationships or features that influence their accuracy are often hard to interpret. Here, we developed DLFea4AMPGen as a strategy for de novo AMP generation that utilizes AA features learned from deep learning models. Based on SHAP estimates of “black box” feature influence in deep learning models, our strategy simultaneously considers antibacterial, antifungal, and antioxidant activities within a comprehensive peptide feature space, ensuring that the de novo AMPs incorporate features relevant to all of these target activities, ultimately increasing the success rate. Experimental validation showed that 12 of the 16 c-AMPs in our current study exhibited bioactivity, resulting in a positive rate of 75%, thus providing a proof-of-concept demonstration that DLFea4AMPGen has a robust capacity for generating candidate AMPs.

Unlike previous studies that predict all sequences, we constructed comprehensive peptide sequence spaces with a high positive rate based on the model's strong feature extraction capabilities, which could reduce the screening scale while ensuring the accuracy of the entire strategy. In our study, by extracting the AAs that have the greatest influence on bioactivity to construct the plausible sequence subspace, this strategy ensures that the generated AMPs indeed contain the features essential for antimicrobial activity. Most importantly, by focusing on the most significant features in SHAP analysis rather than random combinations of 20 AAs, to some extent, it addresses the global search problem for peptide exploration²³. In addition, compared to more common database filtering techniques that yield random peptide combinations comprised of only a few (e.g., three) AAs^{24,25,37}, we conducted a systematic analysis of all sequences with potential antibacterial activity and selected representative sequences from among groups of similar sequences to maximize the diversity of the de novo c-AMPs.

AMPs have garnered significant attention in recent research. Multiple studies have highlighted the impact of factors like charge, secondary structure, hydrophobicity, and amphiphilicity on the substrate specificity and bioactivity of peptides in disrupting bacterial cell membranes^{8,30,38–40}. Obviously, these 16 de novo c-AMPs possess these basic characteristics. Using the traditional motif extraction method (MEME)⁴¹, we combined the three datasets of ABP, AFP, and AOP to identify motifs. However, the extracted motifs did not exhibit clear patterns. This discrepancy may be due to the fact that, unlike enzymes, antimicrobial peptides are shorter in length and do not rely on an active site for their function. Therefore, the motifs extracted by MEME

may not capture the general functional mechanisms of AMPs, which limits their utility for de novo AMP design. In this study, DLFea4AMPGen allowed us to consider more than one activity type in our AMP design agenda by integrating multiple features into a single sequence, resulting in peptides with strong antibacterial, antifungal, and antioxidant activities. In addition to antibacterial activity, increasing the prevalence of Trp, Cys, and Pro, which are crucial for antioxidant activity^{42,43}, could help reduce oxidative stress caused by bacterial infections and effectively decrease host inflammatory response^{26,44,45}.

In this study, the unique mode of action mediated by D1, which involves the destruction of cell membrane structures, resulted in broad-spectrum antibacterial activity against 17 microorganisms, including antibiotic-resistant *S. aureus* strains that do not respond to clinical antibiotic treatments. Furthermore, we observed that D1 not only eliminated bacteria but also lowered inflammatory factor levels, which could help improve treatment outcomes and prevent some complications^{46,47}. However, despite the high efficacy of AMPs produced by our strategy in the current study, our predictive models could be further optimized in future work. For instance, a regression model for MIC prediction could facilitate quantitative evaluation of AMP bioactivity²³, rather than providing only a qualitative assessment. As reported previously⁹, a pipeline composed of multiple models can substantially improve prediction accuracy, reducing the cost of experimental validation. In addition, incorporating an MIC prediction model and empirical pre-screening²³ could further refine the selection process based on quantitative results, enabling a more precise narrowing of the range and ultimately enhancing the accuracy of the final experimental validation.

Here, we introduce DLFea4AMPGen, a deep learning approach for AMP generation that learns features associated with peptide bioactivity. DLFea4AMPGen can reduce computational costs, enhance design success rates, and produce AMPs with more diverse activities compared with other current methods for AMP design. Through SHAP analysis, we identified the features and physical factors with the strongest influence on prediction accuracy in the deep learning models, then focused specifically on these features to construct a plausible sequence subspace. Using this sequence space to design de novo AMPs, we synthesized 16 AMPs for experimental validation, and detected robust, broad-spectrum antimicrobial activity, including antibiotic-resistant strains, by the candidate AMPs, D1 and D2. The efficacy of D1 and D2 provides proof-of-concept evidence validating our approach and supporting the further development of therapeutics targeting drug-resistant pathogens. DLFea4AMPGen can streamline AMP discovery and pave the way for future advances in peptide design applications.

Methods

Ethics statement

All experimental procedures were conducted in accordance with the guidelines of the Institute of Animal Science of the Chinese Academy of Agricultural Sciences. The study protocol was approved by the Institutional Animal Care and Ethics Committee (approval no. IAS2024-116). This study adhered to ethical guidelines for animal research.

Model establishment and evaluation

All 20 benchmark datasets used as predicted objects in this study are listed in Supplementary Table 1, which were sourced from a previous study³². Building on the theory of transfer learning, we fine-tuned all the hidden layers of the pre-trained MP-BERT³³ using the ABP, AFP, and AOP datasets to construct the models, resulting in ABP-MPB, AFP-MPB, and AOP-MPB. In the input layer, AAs were treated as individual text units, separated by spaces, with sequences marked by “CLS” at the start and “SEP” at the end. Sequence vectors were padded with 0 if the

raw sequence was shorter than 1,024 AAs. BAP-MPB incorporates a six-layer hidden layer structure, with initialization weights for the hidden layers derived from the pre-trained model. The final output layer is a fully connected network that maps the learned peptide representations to class labels.

All positive peptides in the 20 datasets from the previous study were used as predicted objects, with ABP, AFP, and AOP used to construct the models for the strategy of DLFea4AMPGen. In order to compare the performance with existing large language models, we further built 20 models for all 20 bioactive peptide datasets separately. To maintain consistency with the UniDL4Biopep³² structure, we replaced the output layer of MP-BERT with the same CNN network as UniDL4Biopep, rather than fine-tuning it. This model is named BERT + CNN. Then, we trained three models (ESM + CNN, BERT + CNN, and BAP-MPB) using the same dataset partitioning method in 20 bioactive datasets, respectively. Model performance was evaluated using the area under the curve, accuracy, and Matthews correlation coefficient, calculated with the Python package scikit-learn (version 1.0.2). To prevent overfitting, we employed an ‘early-stop’ strategy, setting a maximum of 200 training epochs and an early stopping criterion with a patience of 50 rounds. Training was stopped and the model saved if there was no improvement in validation performance within 50 consecutive epochs. Training was conducted using the Mindspore framework (version 1.8).

To ensure fair and unbiased performance evaluation and comparison, we adopted the same dataset division method used in the original research. In particular, the original test datasets from previous work remained unchanged and were not included in model training. Additionally, the original training datasets were categorized into 10 parts for 10-fold cross-validation. Each part was sequentially used as a validation dataset, whereas the remaining nine parts served as the new training dataset for each fold. The average predictive performance on the test dataset during the 10-fold cross-validation was used as the final evaluation criterion. Further details about the benchmark datasets can be found in the referenced previous study.

Analysis of amino acid feature attribution using SHAP

To design a bioactive peptide with antibacterial, antifungal, and antioxidant activities, we calculate the SHAP values for each AA using the ABP-MPB, AFP-MPB, and AOP-MPB models, respectively. Due to significant threshold differences in SHAP values from different models, we first normalized the SHAP values of all AAs in each sequence. This normalization explains the relative contribution of each AA. Next, we calculated the average normalized SHAP values from the three models to serve as the final standard for evaluating the contribution of each AA. Thus, the average normalized SHAP value from the three models was a crucial selection criterion, incorporating features of all three activities.

The SHAP method is an advanced machine learning tool used to interpret the outputs of predictive models⁴⁸. For the opaque operations of complex deep learning models, SHAP provides a framework to understand the specific contributions of each feature to model predictions. For a given sequence and its activity prediction, each AA is associated with a SHAP value. SHAP can explain individual predictions, demonstrating the influence of individual features on the final outcome. By deep learning models, we applied the Python package SHAP (version 0.41.0) to calculate the contribution of each AA in each sequence. The predicted score for each sequence is the sum of the base value and the SHAP values of each AA. By analyzing these SHAP values, we can identify the potential positive or negative impacts of AAs on the prediction score as well as the extent of these impacts.

Key feature fragments containing three activities

For all positive samples in the 20 bioactive peptide datasets, we focused on those predicted as positive in 10-fold cross-validation by the ABP-MPB, AFP-MPB, and AOP-MPB models simultaneously. Among

these peptides with potential triple activities, a length of 13 AAs was found to be the most common. Consequently, a length of 13 AAs was selected for KFF extraction and AMP design.

First, we set a sliding window of 13 AAs with a step size of 1 AA and slid it across the entire sequence. In each sequence, the average SHAP values of each AA served as the basic evaluation metric. In particular, we retained only one window with the highest sum of average SHAP values in each sequence, provided the sum was greater than 0, ensuring the selection of the most distinctive features. Subsequently, an evolutionary relationship analysis was conducted on all KFFs obtained from the focused sequences. Multiple sequence alignment was performed using the maximum likelihood method with MEGA11 software. The JTT + CAT model from FastTree software (version 2.0.0) was used to construct the phylogenetic tree, grouping KFFs with similar AA features into the same subfamily. Finally, the phylogenetic tree was visualized using the Chiplot online website (<https://www.chiplot.online>)⁴⁹.

Constitution of the entire scanning space and de novo design of AMPs

For each subfamily with significant features, we constructed a distinct scanning space. To emphasize the most significant AA features and reduce computational costs, we retained only the top three AAs with the highest frequency at each of the 13 positions. In particular, after aligning all KFFs of the same length within each subfamily, we calculated the frequency of each AA at each position. Then, the top three most frequent AAs at each position were used to construct the plausible sequence subspace. Consequently, each distinct scanning space contained 3^{13} (1,594,323) candidate sequences, each based on the unique key features of its respective subfamily.

Subsequently, during the 10-fold cross-validation of the ABP-MPB, AFP-MPB, and AOP-MPB models, the best-performing model in terms of prediction accuracy was used to filter out sequences that are least likely to be active. More stringent screening criteria were further applied to enhance the overall success rate of the strategy: sequences with model prediction scores below 99% were excluded, and sequences with fewer than two positive charges were disregarded based on empirical rules³⁶.

Among sequences that fulfilled all the aforementioned criteria, representative sequences were selected for further validation. This process was primarily conducted simultaneously across all subfamilies using iFeatureOmegaCLI⁵⁰ in Python. First, Amphiphilic Pseudo AA Composition (APAAC) was used to characterize the feature vectors of sequences. By incorporating pseudo AA components, APAAC combines AA properties (such as hydrophilicity and hydrophobicity) and sequence order information to generate biologically meaningful feature descriptors. Next, principal component analysis was applied to reduce the raw dimensional data generated by APAAC to two dimensions, thereby reducing data complexity. Then, K-means clustering was applied, experimenting with different cluster numbers ranging from 2 to 9. The `kelbow_visualizer` function from the Yellowbrick package was used to determine the optimal cluster number based on the elbow method. Finally, sequences closest to each cluster center were selected as the representative sequences and chemically synthesized for functional validation. And all peptides with purity >95% were custom-synthesized by GenScript Co., Ltd (Nanjing, China).

Structure prediction

The tertiary structures of representative peptides were predicted using ColabFold, employing the `alphafold2_ptm` model by default to generate five model variants. The confidence in these models is indicated by the predicted local distance difference test (pLDDT) scores, which range from 0 to 100. We selected the model with the highest pLDDT score for further analysis and visualized it using PyMOL (version 2.4).

Sequence similarity estimation

We applied the Needleman-Wunsch algorithm in the function `global alignment` from the Biopython library (version 1.76) to estimate the similarity between our query peptide and protein sequences in the dataset, which were ADAM, APD, CAMPR4, DBAASP, and BIOPEP. The alignment was performed using default parameters, and the similarity was calculated by dividing the alignment score by the maximum length of the sequences being compared.

Bacterial strains and growth conditions

Bacterial strains utilized in this work are listed in Supplementary Table 3. Briefly, six bacterial strains were preserved in the laboratory, which were *E. coli* O157:H7, *K. pneumoniae* ATCC700603, *S. typhimurium* ATCC14028, *M. luteus* ATCC4698, *B. subtilis* WB600, and *S. aureus* ATCC6538. Four drug-resistant bacterial strains were isolated from diseased chickens on livestock, which were *E. coli* z44, *E. coli* z59, *S. typhi* MJ1, and *S. typhi* XYJ1. Another six drug-resistant bacterial strains were isolated clinically from patients with various infections in the Chinese PLA General Hospital (301 Hospital), which were *S. aureus* 09057, *S. aureus* 11533, *S. aureus* 12513, *S. aureus* 12533, *S. aureus* 18437, and *S. aureus* 23156.

Antibacterial activity assay

Antibacterial activity was assessed according to the Clinical and Laboratory Standards Institute guidelines³¹. Initially, 16 cAMPs were screened at a concentration of 128 μ M for their inhibitory activity against three Gram-negative bacteria (*E. coli* O157:H7, *K. pneumoniae* ATCC700603, and *S. typhimurium* ATCC14028) and three Gram-positive bacteria (*M. luteus* ATCC4698, *B. subtilis* WB600, and *S. aureus* ATCC6538). Bacteria were inoculated in cation-adjusted Mueller-Hinton broth (CaMHB; no. HB6231-I; Qingdao Hope Bio-Technology Co., Ltd., Shandong, China) and incubated at 37 °C overnight. Then, the cultures were diluted (1:100) with fresh CaMHB and cultured to the exponential phase (OD₆₀₀ of 0.4–0.6). The cell concentration was adjusted to approximately 5×10^5 c.f.u. per mL. Next, 100- μ L aliquots were transferred into 96-well plates containing 100 μ L of different cAMP solutions. For MIC determination, the tested ranges of AMPs were serially diluted two-fold from 128 μ M to 1 μ M. Bacterial growth was monitored using the automated Bioscreen C microbiological growth monitoring system (LabSystems, Helsinki, Finland). The system was incubated at 37 °C for 16 h, with OD₆₀₀ measurements obtained every 10 min. By plotting the growth curves, we observed the antibacterial activities of the peptides on different bacteria. The bacteriostatic rate was calculated using the following equation⁵²:

$$\text{Inhibition rate (\%)} = 1 - \frac{A_s - A_{s'}}{A_b - A_{b'}} \times 100 \quad (1)$$

where A_s is the absorbance of the AMP treatment group, A_b is the absorbance of the control group, $A_{s'}$ is the initial absorbance of the AMP treatment group, and $A_{b'}$ is the initial absorbance of the control group.

Antifungal activity assay

Similarly, at a concentration of 128 μ M, 16 cAMPs were initially screened for their inhibitory activity against the fungus *C. albicans* ATCC10231. The experimental procedure was similar to that for bacterial inhibition, with some differences: *C. albicans* ATCC10231 was cultured in potato dextrose broth at an incubation temperature of 30 °C, with the duration of the reaction doubled compared to the antibacterial activity assay.

Cell culture and cytotoxicity assays

NIH/3T3 cells (a murine fibroblast cell line) were obtained from American Type Culture Collections (ATCC CRL-1658). For the

cytotoxicity test, 3T3 cells were seeded into 96-well plates, with a volume of 100 μ L per well. The cells were incubated at 37 °C with 5% CO₂ for approximately 18 h. Once the cells had adhered and reached approximately 40% confluence, the culture medium was replaced with 100 μ L of fresh medium containing 128 μ M of the test compound ($n = 3$). The cells were then cultured for an additional 24 h under the same conditions. After the 24-h incubation, 10 μ L of CCK-8 reagent was added to each well, and the plates were incubated for 1.5 h at 37 °C. The absorbance at 450 nm was measured using a microplate reader. Each experimental condition was performed in quintuplicate. Triton X-100 was used as the positive control. The percentage of cell inhibition was calculated using the following formula:

$$\% \text{Inhibition} = \frac{\text{absorbance}_{\text{negative}} - \text{absorbance}_{\text{sample}}}{\text{absorbance}_{\text{negative}} - \text{absorbance}_{\text{blank}}} \times 100 \quad (2)$$

Hemolysis test

Red blood cells (RBCs) from rabbits (no. S27562-100ml; Shanghai Yuanye Bio-Technology Co., Ltd, Shanghai, China) were used to assess the hemolytic activity of peptides. Peptide solutions ($n = 3$) were mixed with RBCs to achieve a final concentration of 128 μ M. The mixtures were incubated at 37 °C for 4 h. RBCs in phosphate-buffered saline (PBS) served as the negative control, whereas RBCs in H₂O served as the positive control. After incubation, the samples were centrifuged at $2500 \times g$ for 10 min, and the absorbance of the supernatant was measured at 540 nm using a microplate reader. The percentage of hemolysis was calculated as:

$$\% \text{Hemolysis} = \frac{\text{absorbance}_{\text{sample}} - \text{absorbance}_{\text{negative}}}{\text{absorbance}_{\text{positive}} - \text{absorbance}_{\text{negative}}} \times 100 \quad (3)$$

Drug resistance assay

Resistance development assays were performed following a previous report²³. Drug resistance was tested by repeatedly treating *S. aureus* 09057 with peptide D1 at the sub-MIC concentration (1/2 MIC). MIC values were tested using the described method after every 10 generations. All experiments were performed with three independent replicates.

Antioxidant activity assay

The antioxidant activity of cAMPs was evaluated by measuring their ability to scavenge ABTS⁺ free radicals at a concentration of 1 mg/mL. Glutathione at the same concentration was used as the positive control, and double-distilled water served as the negative control.

ABTS⁺ radical scavenging assay: We weighed 200 mg ABTS and 34.4 mg potassium persulfate, dissolved them in 50.0 ml of double-distilled water, mixed the solution thoroughly, and let it stand at room temperature in the dark for 24 h. Then, the ABTS assay solution was prepared by diluting an appropriate quantity of the ABTS stock solution with 95% ethanol until the absorbance value reached 0.70 ± 0.02 (OD₇₃₄). Then, 180 μ L aliquots of the assay solution were transferred to 96-well plates containing 20 μ L of different cAMP solutions. After thoroughly mixing each sample, the reaction was allowed to proceed at room temperature in the dark for 5 min. The absorbance was measured at a wavelength of 734 nm, using the sample solvent for zero calibration.

The free radical scavenging rate was calculated as follows:

$$\% \text{Scavenging rate} = \left(\frac{A_b - A_s}{A_b} \right) \times 100 \quad (4)$$

where A_b is the absorbance of the mixture of ABTS solution and the sample solvent solution, A_s is the absorbance of the mixture of the test solution and ABTS solution.

SEM and TEM measurement

Membrane permeability induced by AMPs was measured using SEM and TEM. *S. aureus* 09057 was grown overnight at 37 °C. Bacterial suspensions (c.f.u. = 10^8 ml⁻¹) were treated with D1 at final concentrations of 64 μM at 37 °C for 2 h and 3 h, with untreated cells serving as the control. Samples were washed with 0.1 M PBS three times and treated with glutaraldehyde for 4 h. A 10 μL sample was added onto a silicon wafer and sequentially dehydrated with 20%, 40%, 60%, 80%, and 100% ethanol/water (10 min each). The specimens were observed using SEM (Hitachi SU3500, Japan). For ultrathin sectioning, samples were sequentially treated with glutaraldehyde overnight and osmic acid (1 wt%) for 2 h. Then they were dehydrated with 30%, 50%, 70%, 80%, 90%, and 100% ethanol/water (30 min each) and acetone (20 min). Subsequently, the specimens were infiltrated in a series of acetone and Spurr's resin mixtures (1:1 for 1 h, 1:3 for 3 h) and finally embedded in Spurr's resin overnight. Ultrathin sections of about 70 nm thickness were stained with 2% uranyl acetate and lead citrate (10 min each). Images were taken with a TEM (Hitachi HT7700, Japan).

In vivo experiments

We established the mouse sepsis model²⁹ using *S. aureus* 09057 and *E. coli* z44 under identical experimental conditions, respectively. Five-week-old healthy female BALB/c mice (inbred, wild-type) were purchased from SPF (Beijing) Biotechnology Co., Ltd. and randomly divided into four groups, with six mice per group. They were acclimated for 7 days in a new environment with suitable conditions: a temperature range of 20–24 °C, humidity levels of 50–60%, and alternating periods of light and darkness. Experiments were performed when the mice reached 6 weeks of age, at which point their body weight was approximately 20 g. Mice in both the negative control group and the treatment group were intraperitoneally injected with bacterial suspension at a dose of 10^8 CFU, whereas the blank group was injected with 150 μL of physiological saline. After 1 h, the negative control group was intraperitoneally injected with 200 μL physiological saline, the treatment group was intraperitoneally injected with 200 μL D1 (20 mg/kg), and the blank group received no treatment. After 12 h, all mice were euthanized using ether inhalation anesthesia. The liver, spleen, lungs, and kidneys of each mouse were collected, weighed, and homogenized. The homogenates were serially diluted in physiological saline, plated onto solid LB, and the c.f.u. of bacteria per organ was calculated by counting colonies on agar plates. Results were expressed as CFU/g of tissue for each organ. On the other hand, blood was collected from the orbital vein of the mice, and serum levels of cytokines were analyzed using an enzyme-linked immunoassay, which were TNF-α, IL-1β, and IL-6. The inflammatory factor kit was purchased from Beijing Solarbio Science & Technology Co., Ltd. (Beijing, China). Statistical analysis was conducted using one-way ANOVA.

Statistics and reproducibility

All experiments were independently repeated at least twice with similar results to ensure reproducibility. Sample sizes for each experiment were determined based on previous published studies in the field, ensuring sufficient statistical power. For in vitro assays, three biologically independent replicates were used per condition. For animal studies, six biologically independent mice were used per group. Animals were randomly assigned to different treatment groups. Blinding was not performed due to the nature of the treatment and infection procedures. Nonetheless, no apparent sources of bias were identified, and data collection was carried out using standardized protocols and objective measurements to ensure reliability. No

statistical method was used to predetermine sample size. No data were excluded from the analyses.

Reporting summary

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

Data availability

The 20 datasets used for model construction were obtained from the previous study and are publicly available at GitHub: <https://github.com/dzjxzyd/UniDL4BioPep>. The datasets used for sequence similarity comparison were uploaded to GitHub and were derived from five public databases: ADAM, APD3 (<http://aps.unmc.edu/>), CAMPR4 (<http://camp.bicnirrh.res.in/>), DBAASP, and BIOPEP. The curated similarity datasets are available at: https://github.com/hgao12345/DLFea4AMPGen/blob/main/Dataset/AMP_from_5databases.xlsx. The trained model checkpoint is available on Zenodo (<https://zenodo.org/records/16545412>). All source data supporting the findings of this study are available within the main text, Supplementary Information, or via the provided repositories. Source data are provided with this paper.

Code availability

The models and source codes used for DLFea4AMPGen are available at GitHub (<https://github.com/hgao12345/DLFea4AMPGen>) and archived on Zenodo with the identifier [<https://doi.org/10.5281/zenodo.16681867>]³³. Moreover, Pre-trained MP-BERT can be accessed at <https://github.com/BRITian/MP-BERT>.

References

- Deo, S., Turton, K. L., Kainth, T., Kumar, A. & Wieden, H.-J. Strategies for improving antimicrobial peptide production. *Biotechnol. Adv.* **59**, 107968 (2022).
- Ngambenjawong, C., Chan, L. W., Fleming, H. E. & Bhatia, S. N. Conditional antimicrobial peptide therapeutics. *ACS Nano* **16**, 15779–15791 (2022).
- Magana, M. et al. The value of antimicrobial peptides in the age of resistance. *Lancet Infect. Dis.* **20**, e216–e230 (2020).
- Fjell, C. D., Hiss, J. A., Hancock, R. E. & Schneider, G. Designing antimicrobial peptides: form follows function. *Nat. Rev. Drug Discov.* **11**, 37–51 (2012).
- Lazzaro, B. P., Zasloff, M. & Rolff, J. Antimicrobial peptides: application informed by evolution. *Science* **368**, eaau5480 (2020).
- Zhang, L. -j & Gallo, R. L. Antimicrobial peptides. *Curr. Biol.* **26**, R14–R19 (2016).
- Nguyen, L. T., Haney, E. F. & Vogel, H. J. The expanding scope of antimicrobial peptide structures and their modes of action. *Trends Biotechnol.* **29**, 464–472 (2011).
- Yeaman, M. R. & Yount, N. Y. Mechanisms of antimicrobial peptide action and resistance. *Pharmacol. Rev.* **55**, 27–55 (2003).
- Ma, Y. et al. Identification of antimicrobial peptides from the human gut microbiome using deep learning. *Nat. Biotechnol.* **40**, 921–931 (2022).
- Wang, J. Seeking old wisdoms for new AMP discovery. *Cell Host Microbe* **31**, 1251–1253 (2023).
- Santos-Júnior, C. D. et al. Discovery of antimicrobial peptides in the global microbiome with machine learning. *Cell* **187**, 3761–3778 (2024).
- Wan, F., Torres, M. D., Peng, J. & de la Fuente-Nunez, C. Deep-learning-enabled antibiotic discovery through molecular de-extinction. *Nat. Biomed. Eng.* **8**, 854–871 (2024).
- Guo, X. et al. Novel Feleucin-K3-derived peptides modified with sulfono-γ-AA building blocks targeting *Pseudomonas aeruginosa* and methicillin-resistant *Staphylococcus aureus* infections. *J. Med. Chem.* **66**, 1254–1272 (2022).

14. Zhang, H. et al. Machine learning and genetic Algorithm-Guided directed evolution for the development of antimicrobial peptides. *J. Adv. Res.* **68**, 415–428 (2025).
15. Wan, F., Wong, F., Collins, J. J. & de la Fuente-Nunez, C. Machine learning for antimicrobial peptide identification and design. *Nat. Rev. Bioeng.* **2**, 392–407 (2024).
16. Xu, J. et al. iAMPNCN: a deep-learning approach for identifying antimicrobial peptides and their functional activities. *Brief. Bioinform.* **24**, bbad240 (2023).
17. Goverde, C. A. et al. Computational design of soluble and functional membrane protein analogues. *Nature* **631**, 449–458 (2024).
18. Hsu, C., Fannjiang, C. & Listgarten, J. Generative models for protein structures and sequences. *Nat. Biotechnol.* **42**, 196–199 (2024).
19. Pandi, A. et al. Cell-free biosynthesis combined with deep learning accelerates de novo-development of antimicrobial peptides. *Nat. Commun.* **14**, 7197 (2023).
20. Das, P. et al. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nat. Biomed. Eng.* **5**, 613–623 (2021).
21. Ingraham, J. B. et al. Illuminating protein space with a programmable generative model. *Nature* **623**, 1070–1078 (2023).
22. Madani, A. et al. Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **41**, 1099–1106 (2023).
23. 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).
24. Xu, B. et al. Gut-targeted nanoparticles deliver specifically targeted antimicrobial peptides against *Clostridium perfringens* infections. *Sci. Adv.* **9**, eadf8782 (2023).
25. Li, J.-x. et al. De novo design and antibacterial activity of α -helical antimicrobial peptide YHX-1. *J. Food Sci. Technol.* **40**, 98–104 (2022).
26. Bhattacharyya, A., Chattopadhyay, R., Mitra, S. & Crowe, S. E. Oxidative stress: an essential factor in the pathogenesis of gastrointestinal mucosal diseases. *Physiol. Rev.* **94**, 329–354 (2014).
27. Si, M. et al. Manganese scavenging and oxidative stress response mediated by type VI secretion system in *Burkholderia thailandensis*. *Proc. Natl. Acad. Sci. USA* **114**, E2233–E2242 (2017).
28. Allaoui, A., Botteaux, A., Dumont, J. E., Hoste, C. & De Deken, X. Dual oxidases and hydrogen peroxide in a complex dialogue between host mucosae and bacteria. *Trends Mol. Med.* **15**, 571–579 (2009).
29. Tan, P. et al. Designing self-assembling chimeric peptide nanoparticles with high stability for combating piglet bacterial infections. *Adv. Sci.* **9**, 2105955 (2022).
30. Li, Y. et al. A dual functional polypeptide with antibacterial and anti-inflammatory properties for the treatment of periodontitis. *Int. J. Biol. Macromol.* **242**, 124920 (2023).
31. Dong, R. et al. Exploring the repository of de novo-designed bifunctional antimicrobial peptides through deep learning. *eLife* **13**, RP97330 (2025).
32. Du, Z., Ding, X., Xu, Y. & Li, Y. UniDL4BioPep: a universal deep learning architecture for binary classification in peptide bioactivity. *Brief. Bioinform.* <https://doi.org/10.1093/bib/bbad135> (2023).
33. Liu, T. et al. Protein–protein interaction and site prediction using transfer learning. *Brief. Bioinform.* **24**, bbad376 (2023).
34. Cao, Q. et al. Designing antimicrobial peptides using deep learning and molecular dynamic simulations. *Brief. Bioinform.* **24**, bbad058 (2023).
35. Lundberg, S. M. & Lee, S.-I. A unified approach to interpreting model predictions. *Adv. Neural Inf. Process. Syst.* **30**, 4768–4777 (2017).
36. Chen, N. & Jiang, C. Antimicrobial peptides: structure, mechanism, and modification. *Eur. J. Med. Chem.* **255**, 115377 (2023).
37. Mishra, B. & Wang, G. Ab initio design of potent anti-MRSA peptides based on database filtering technology. *J. Am. Chem. Soc.* **134**, 12426–12429 (2012).
38. Mahlapuu, M., Håkansson, J., Ringstad, L. & Björn, C. Antimicrobial peptides: an emerging category of therapeutic agents. *Front. Cell. Infect. Microbiol.* **6**, 235805 (2016).
39. Li, S. et al. The structure-mechanism relationship and mode of actions of antimicrobial peptides: a review. *Trends Food Sci. Technol.* **109**, 103–115 (2021).
40. Shastri, J. et al. Effect of dietary supplementation of Antimicrobial peptide on production performance, egg quality and serum biochemical parameters of laying hens. *Asian Journal of Dairy and Food Research* **44**, 326–331 (2025).
41. Bailey, T. L., Johnson, J., Grant, C. E. & Noble, W. S. The MEME suite. *Nucleic Acids Res.* **43**, W39–W49 (2015).
42. Elias, R. J., Kellerby, S. S. & Decker, E. A. Antioxidant activity of proteins and peptides. *Crit. Rev. Food Sci. Nutr.* **48**, 430–441 (2008).
43. Zheng, L., Zhao, Y., Dong, H., Su, G. & Zhao, M. Structure–activity relationship of antioxidant dipeptides: dominant role of Tyr, Trp, Cys and Met residues. *J. Funct. Foods* **21**, 485–496 (2016).
44. Zhu, Y. et al. Antioxidant peptides, the guardian of life from oxidative stress. *Med. Res. Rev.* **44**, 275–364 (2024).
45. Tonolo, F. et al. Sunflower seed-derived bioactive peptides show antioxidant and anti-inflammatory activity: From in silico simulation to the animal model. *Food Chem.* **439**, 138124 (2024).
46. Zainel, A., Mitchell, H. & Sadarangani, M. Bacterial meningitis in children: neurological complications, associated risk factors, and prevention. *Microorganisms* **9**, 535 (2021).
47. Grabež, M. et al. A prospective, randomized, double-blind, placebo-controlled trial of polyphenols on the outcomes of inflammatory factors and oxidative stress in patients with type 2 diabetes mellitus. *Rev. Cardiovasc. Med.* **23**, 57 (2022).
48. Rozemberczki, B. et al. The Shapley value in machine learning. *Proc. 31st Int. Joint Conf. Artif. Intell.* 5572–5579 (2022).
49. Xie, J. et al. Tree Visualization By One Table (tvBOT): a web application for visualizing, modifying and annotating phylogenetic trees. *Nucleic Acids Res.* **51**, W587–W592 (2023).
50. Chen, Z. et al. iFeatureOmega: an integrative platform for engineering, visualization and analysis of features from molecular sequences, structural and ligand data sets. *Nucleic Acids Res.* **50**, W434–W447 (2022).
51. Humphries, R., Bobenchik, A. M., Hindler, J. A. & Schuetz, A. N. Overview of changes to the clinical and laboratory standards institute performance standards for antimicrobial susceptibility testing, M100. *J. Clin. Microbiol.* **59**, e0021321 (2021).
52. Yu, R. et al. Engineering antimicrobial metal–phenolic network nanoparticles with high biocompatibility for wound healing. *Adv. Mater.* **36**, 2307680 (2024).
53. Gao, H. DLFea4AMPGen de novo design of antimicrobial peptides by integrating features learned from deep learning models (Version v1). Zenodo <https://doi.org/10.5281/zenodo.16681867> (2025).

Acknowledgements

This work was financially supported by the National Key R&D Program of China (2022YFD1300700 to B.Y.), the National Natural Science Foundation of China (32130101 to H.L., 32270813 to Y.T., and 32472945 to F.G.), the Agricultural Science and Technology Innovation Program (CAAS-ZDRW202304 to H.L.), and the China Agriculture Research System of MOF and MARA (CARS-41 to Y.W.). We thank Dr. Yong Zhao from the Institute of Animal Science, Chinese Academy of Agricultural Sciences, for his assistance in the operation and data analysis of in vivo mouse experiments.

Author contributions

Han Gao and Feifei Guan: Performed the experiments, analyzed the data and wrote the paper; Boyu Luo: Performed the experiments, analyzed the data; Dongdong Zhang and Wei Liu: Analyzed the data and contributed materials/analysis tools; Yuying Shen, Lingxi Fan, and Guoshun Xu: Performed the experiments and analyzed the data; Yuan Wang, Tao Tu, Ningfeng Wu, and Bin Yao: Analyzed the data; Huiying Luo: Conceived and designed the experiments and analyzed the data; Yue Teng: Conceived and designed the experiments and contributed materials/analysis tools; Jian Tian: Conceived and designed the experiments and analyzed the data; Huoqing Huang: Conceived and designed the experiments, analyzed the data, contributed materials/analysis tools and wrote the paper.

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/s41467-025-64378-y>.

Correspondence and requests for materials should be addressed to Huiying Luo, Yue Teng, Jian Tian or Huoqing Huang.

Peer review information *Nature Communications* thanks the 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.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025