Briefings in Bioinformatics, 2023, 1–13

https://doi.org/10.1093/bib/bbad058

Problem Solving Protocol

Designing antimicrobial peptides using deep learning and molecular dynamic simulations

Qiushi Cao†, Cheng Ge†, Xuejie Wang†, Peta J. Harvey, Zixuan Zhang, Yuan Ma, Xianghong Wang, Xinying Jia, Mehdi Mobli, David J. Craik, Tao Jiang, Jinbo Yang, Zhiqiang Wei, Yan Wang, Shan Chang and Rilei Yu

Corresponding authors Yan Wang. E-mail: wangy12@ouc.edu.cn; Shan Chang. E-mail: schang@jsut.edu.cn; Rilei Yu, Ocean University of China, 5 Yushan Road, Qingdao 266003, China. Tel./Fax: +86-0532-82032951; E-mail: ryu@ouc.edu.cn

†Qiushi Cao, Cheng Ge and Xuejie Wang contributed equally to this work.

Abstract

With the emergence of multidrug-resistant bacteria, antimicrobial peptides (AMPs) offer promising options for replacing traditional antibiotics to treat bacterial infections, but discovering and designing AMPs using traditional methods is a time-consuming and costly process. Deep learning has been applied to the *de novo* design of AMPs and address AMP classification with high efficiency. In this study, several natural language processing models were combined to design and identify AMPs, i.e. sequence generative adversarial nets, bidirectional encoder representations from transformers and multilayer perceptron. Then, six candidate AMPs were screened by AlphaFold2 structure prediction and molecular dynamic simulations. These peptides show low homology with known AMPs and belong to a novel class of AMPs. After initial bioactivity testing, one of the peptides, A-222, showed inhibition against gram-positive and gram-negative bacteria. The structural analysis of this novel peptide A-222 obtained by nuclear magnetic resonance confirmed the presence of an alpha-helix, which was consistent with the results predicted by AlphaFold2. We then performed a structure–activity relationship study to design a new series of peptide analogs and found that the activities of these analogs could be increased by 4–8-fold against *Stenotrophomonas maltophilia* WH 006 and *Pseudomonas aeruginosa* PAO1. Overall, deep learning shows great potential in accelerating the discovery of novel AMPs and holds promise as an important tool for developing novel AMPs.

Keywords: antimicrobial peptides, deep learning, BERT, molecular dynamic simulations

Introduction

In recent years, the misuse of antibiotics has caused a large number of drug-resistant bacteria to emerge, posing a threat to the security of global public health. Antibiotic-resistant infections cause a substantial clinical and economic burden; as a result, when treating a patient with multidrug-resistant infections, the therapy time is prolonged by 6.4–12.7 days, and the total cost is approximately $18 588–$29 069 [1]. Therefore, it is urgent to develop new antimicrobial therapies. Conventional antibiotics are mostly derived from secondary metabolites of microorganisms, plants and animals; antibiotics are produced during life activities and show antimicrobial activity [2]. Antimicrobial peptides (AMPs), also known as host defense peptides, act as the first line of defense for the host to kill bacterial pathogens [3]. AMPs have attracted much attention because of their low toxicity, broad-spectrum antimicrobial activity and low-level antibacterial resistance [4]. AMPs are less likely to evolve resistance because they show significant pharmacodynamic differences from small-molecule antibiotics [5]. Most AMPs are typically 10–50 amino acids long, with a high percentage of basic amino acids and

Qiushi Cao is a graduate student in Ocean University of China. Her research interests include peptide synthesis and structure-activity relationship study.
Cheng Ge is a graduate student in Ocean University of China. His research interests include bioinformatics and biocomputing.
Xuejie Wang is a graduate student at Ocean University of China. Her research focuses on the development of antibacterial resources for marine bacteria.
Peta Harvey is a senior research fellow at the Institute for Molecular Bioscience, The University of Queensland. Her interest focuses on peptide molecular NMR structure determination.
Zixuan Zhang is a graduate student in Ocean University of China. His research interests include peptide synthesis and bioinformatics.
Yuan Ma is an undergraduate student in Ocean University of China. His research interest is peptide synthesis.
Wang Xianghong is a professor at Ocean University of China. His research interests are marine microbial diversity and their active substances.
Xinying Jia is a senior research fellow at Australian Institute for Bioengineerin and Nanotechnology, The University of Queensland. His research interests include structural and biochemical characterisation of enzymes and biosynthesis of compounds using biocatalysis.
Mehdi Mobli is a professor at Australian Institute for Bioengineerin and Nanotechnology, The University of Queensland. His group focuses on developing methods for biomolecular nmr and using natural peptide ligands to understand ion-channel function.
David J. Craik is a professor at the Institute for Molecular Bioscience, The University of Queensland. His interest focuses on peptide drug discovery and design.
Tao Jiang is a professor in Ocean University of China. Her research interest is pharmaceutical design and synthesis.
Jinbo Yang is a chief professor of "Blue Medicine Talent" of Ocean University of China. His research direction is cancer biology, tumor immunology and molecular pharmacology, and high-throughput drug screening and intelligence supercomputer based virtual drug screening.
Zhiqiang Wei is a professor in Ocean University of China. His research interest is biocomputing.
Shan Chang is a professor at Institute of Bioinformatics and Medical Engineering, School of Electrical and Information Engineering, Jiangsu University of Technology, China. His research interests include bioinformatics, artificial intelligence and molecular simulation.
Yan Wang is a professor at Ocean University of China. His research focuses on the discovery and mechanism of antimicrobial resources of marine microorganisms and bacterial resistance. 16.Rilei Yu is a professor at the School of Medicine and Pharmacy, in the Ocean University of China, China. His research interests are focused on marine peptide drug discovery, peptide structure-activity relationship study, and computer or artificial intelligence aided drug design.

Received: November 7, 2022. **Revised:** January 30, 2023. **Accepted:** February 1, 2023

amphiphilic structures [6]. AMPs function by targeting bacterial cell membranes to form transmembrane ion channels, thereby disrupting the integrity of cell membranes and causing bacterial cell contents to leak, killing pathogenic microorganisms [7]. Recently, some AMPs, such as polymyxin B, tyrothricin and colistins, were approved by the American Food and Drug Administration for clinical use [8].

Artificial intelligence and machine learning have been playing a critical role in pharmaceutical discovery, such as prediction of drug–target binding affinity [9], prediction of potential disease-associated miRNAs [10, 11] and drug design and discovery [12].The *de novo* design of AMPs remains a time-intensive and costly process [3, 13]. Machine learning has also been applied in the design of AMPs, and peptides such as YI12 and FK13 generated using deep-learning classifiers have demonstrated high inhibition potency on the growth of a diverse range of pathogenic microorganisms [14]. Furthermore, the combination of data from the human gut microbiome with neural network models (NNMs) identified a substantial number of bioactive AMPs [2]. Generative adversarial network and variational autoencoder (VAE) methods have been used to study peptide generation [15–17]. Surana *et al.* [15] developed PandoraGAN, which uses a manually curated training data set of 130 highly active peptides that includes peptides from known databases (such as AVPdb) and the literature to generate novel antiviral peptides. Tucs *et al.* [16] presented PepGAN (a peptide-specialized network) to generate a highly active AMP that is twice as strong as ampicillin. Dean *et al.* [17] proposed a model for AMP generation based on VAE. The model sampled from distinct regions of the learned latent space and allowed for controllable generation of new AMP sequences with minimal input parameters. These models were used to generate AMPs without considering the stability of their secondary structures, although that is highly relevant to AMP antimicrobial activity [18].

In the last two decades, a large number of predictors for the activities of AMPs have been developed [2, 19, 20]. First, the predictors are constructed based on conventional methods. Target-AMP for classifying AMPs was designed by Jan *et al.* [21]. The proposed method utilized position-specific scoring matrix, pseudo amino acid composition and dipeptide composition to express peptide sequences. Then, numerous classification techniques, including K-nearest neighbor, random forest (RF) and support vector machine, were used. Second, the first NNM [deep neural network (DNN)]-based classifier proposed by Veltri *et al.* [22] has been applied to achieve better AMP recognition performance. Subsequently, DNN has been used to address AMP classification. The sAMPpredGAT predictor was proposed for the recognition of AMPs by Yan *et al.* [23]. By utilizing structural information, evolutionary profiles and sequence features, the GAT framework extracted the discriminative features from the graph. Finally, the optimized features are fed into the output layer to identify the AMPs. To identify sAMPs, Hussain *et al.* [24] proposed an sAMP-PFPDeep predictor that utilized position, frequency and other physiochemical features through two DNNs, i.e. VGG-16 and RESNET-50. Moreover, deep learning models have demonstrated their ability in various applications [25–27]. Recently, the ESM-2 protein language model was trained on the UniRef50 data set developed by the Facebook team [28]. It is the state-of-the-art protein pretrained language model and is widely used for deep learning tasks based on protein sequence data. In addition, the features learned on millions of sequences are most suitable to perform downstream tasks, such as the binary classification model.

Determining the 3D structures of peptides is vital for understanding their structure–activity relationships (SARs), as well as for the design of more potent analogs [29]. SAR elucidation also plays a key role in facilitating peptide drug development. Experimentally, until the appearance of AlphaFold2, considerable time and energy were needed to obtain the exquisite structures of proteins. AlphaFold2 uses deep learning to predict the structures of proteins with high efficiency and accuracy, which was an outstanding breakthrough in artificial intelligence [30, 31]. Nevertheless, the stability of the structures predicted using AlphaFold2 remains to be validated based on physically based methods, such as molecular dynamics (MD) simulations. Previous experimental studies demonstrated that the functionality of AMPs is highly correlated with the stability of secondary structures, such as helix or beta-strand components [32, 33]; thus, we used the availability of secondary structures as well as peptide stability as an important basis to select the generated peptides.

In this study, we established a flowchart for the efficient design of AMPs by combining deep learning and MD simulations with wet laboratory experimental studies (Figure 1). First, sequence generative adversarial nets (SeqGAN) were used to generate peptides. Then, we proposed an AMP predictor based on bidirectional encoder representations from transformers (BERT) and multilayer perceptron (MLP) to identify SeqGAN-generated peptides. Finally, we applied the following steps to select candidate AMPs: (1) the structures of the candidate AMPs were predicted by AlphaFold2. (2) MD simulations were performed to assess the stability of these peptides. (3) In the wet laboratory, peptides with high stability were synthesized, and their antimicrobial activities were tested.

Materials and methods

Data set collection and preprocessing

AMPs

Most AMPs are typically >10 amino acids in length and the peptides that contain >30 amino acids are not easily synthesized. To create the AMP data set, we collected AMPs of length $\in$ [11, 30] from three public data sets, including ADAM [34], CAMP$_{R4}$ [35] and StarPep [36].

Non-AMPs

The impact of negative data sampling on model performance and benchmarking has recently been developed in detail by Sidorczuk *et al.* [37]. The models mentioned in their study perform better when the training sample is produced by AmpGram [38] (a sampling method that was run on the negative data set of UniProt sequences). Thus, sequences were created by AmpGram of length $\in$ [11, 30] as a non-AMP database in this study.

Preprocessing

With the AMPs and non-AMPs collected, we implemented the following preprocessing steps:

(i) Duplicates and sequences containing unnatural amino acids were removed.
(ii) CD-HIT was adopted to remove redundancy and homologs [39]. The samples in the AMP database were processed by CD-HIT with a threshold of 0.8. Similarly, to avoid overrepresentation of highly similar sequences in non-AMP samples, we adopted CD-HIT to remove sequences with a threshold of 0.4 [39].
(iii) All the negative samples have an equal length distribution with positive samples, as shown in Figure S1.
(iv) To balance the data set, we removed redundant negative samples.

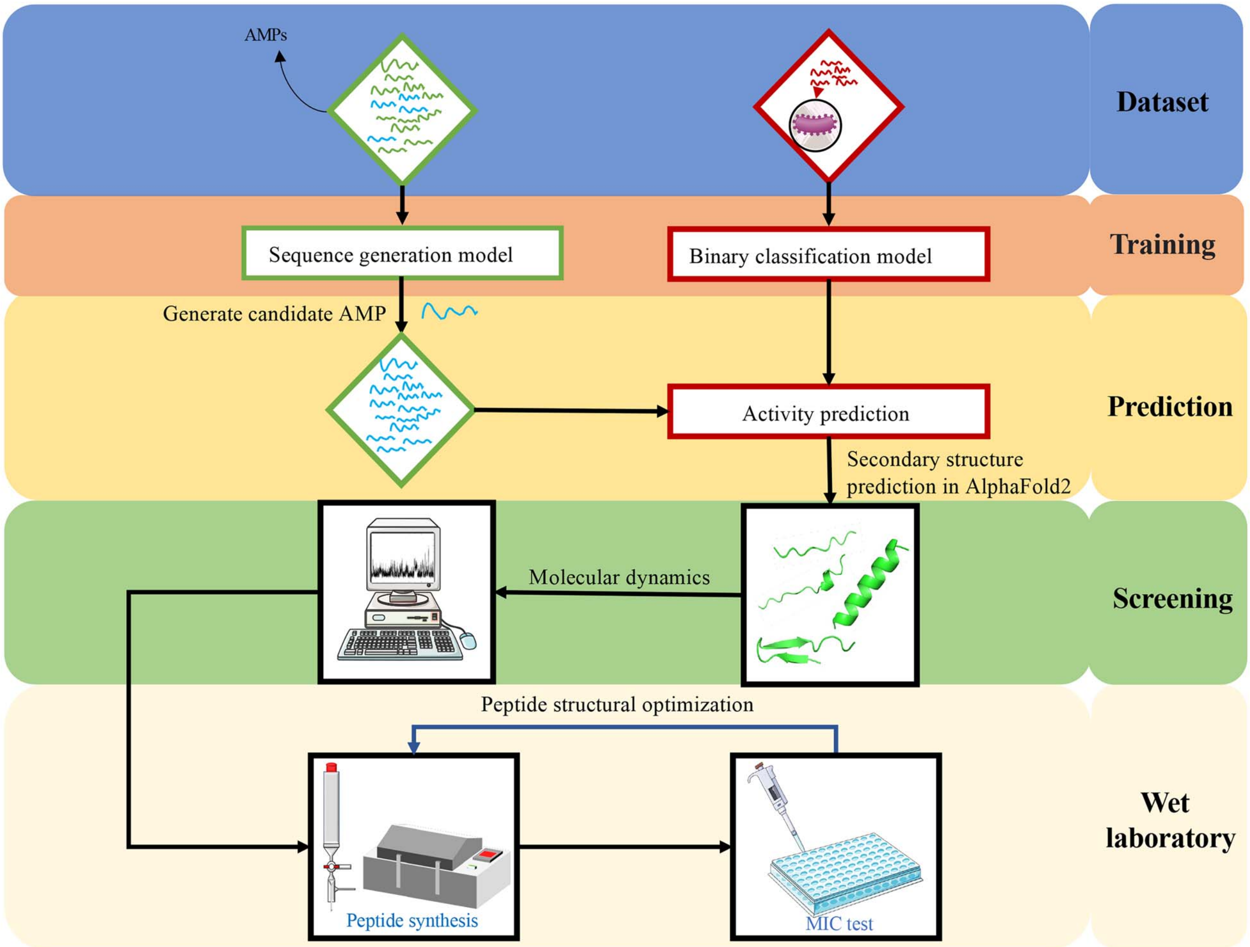

Figure 1. Approach overview. In this study, AMP sequences from the CAMP database were collected and used to train a sequence generation model (left); furthermore, a database was constructed and used to train a binary classification model, which was then applied to identify SeqGAN-generated peptides (right). The screening was completed by combining AlphaFold2 and MD simulations. Chemical synthesis and MIC testing of promising candidates were performed in the wet laboratory. Next, structural modifications were performed to find AMPs with better activity.

Data set splitting

After preprocessing, the final data set comprised 8268 samples (4134 AMPs and 4134 non-AMPs). It was split into a training set and a validation set at a ratio of 8:2. The former is utilized for training the model for classification, and the latter is used for tuning the model's hyperparameters. To objectively evaluate and compare the performance of the proposed model with other models, we constructed an additional independent test data set that contains 162 AMPs that were newly discovered in 2022 from the ADP_3 database [40] and a total of 111 373 non-AMPs from the Sidorczuk *et al.* study [37]. The AMPs in the test data set are not included in the training or validation set, nor are the non-AMPs. Statistical information on the data sets is shown in Table S1.

AMP generation model

In this study, we utilized SeqGAN to generate novel AMPs. We collected 8225 AMPs from the CAMP server [41] and used these AMP sequences as the training set for the AMP generation model. SeqGAN [42] is a text generation method commonly used in natural language processing (NLP) tasks. SeqGAN can be applied to AMP generation tasks by encoding AMP sequences that can be considered text based. The structure of SeqGAN is shown in Figure 2, which consists of the generator, discriminator, Monte Carlo (MC) search and policy gradient. The generator was used to generate the fake peptide, and the discriminator was used to distinguish the real peptide from the fake peptide. We trained the generator by using the MC search and policy gradient. The discriminators were trained by the data of real peptides and generator-generated fake peptides. Through iterative training, the generator produces high-quality candidate peptides that can deceive the discriminator.

AMP classification model

The framework of our classifier is shown in Figure 3. NLP algorithms were applied to construct an AMP classification model. In detail, we incorporated the BERT model and MLP to address this problem. Specifically, the input to BERT was the sequences in the data set, and the output was a vector representation for each sequence. The BERT employed the ESM-2 protein language model, which was trained on the Uniref50 data set [28]. In particular, the features learned on millions of sequences present better classification of AMPs than features learned on small and medium-sized sets [43]. Then, we input the feature representation generated by using BERT into the MLP model to obtain AMP and non-AMP classification results. During the training process, we fixed the BERT model weights and only updated the MLP weights using both AMP and non-AMP data sets. To interpret the proposed model performance in a better way, we used t-distributed stochastic neighbor embedding (t-SNE) [27, 44], which visualizes high-dimensional

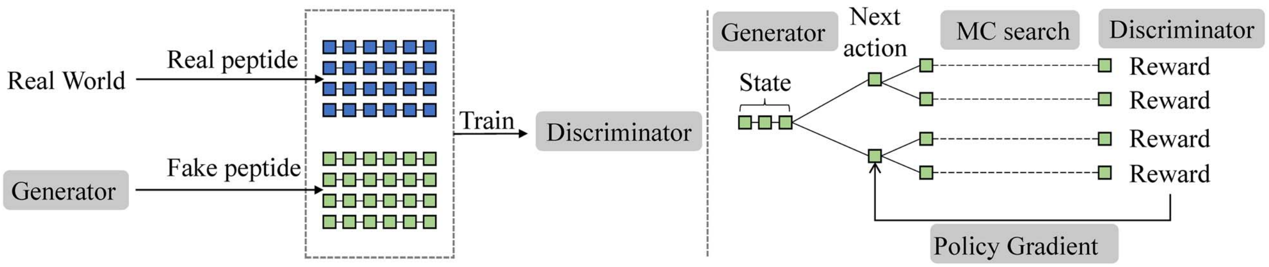

Figure 2. The model structure of SeqGAN. The model utilizes the AMP sequences as input and iteratively trains the generator and discriminator. The trained model outputs candidate peptides.

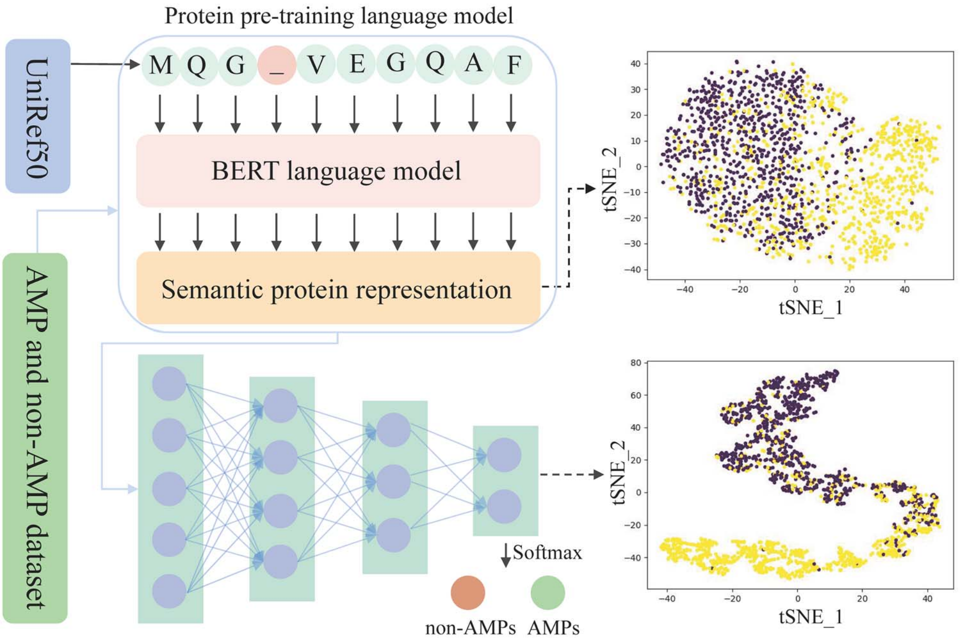

Figure 3. Flowchart of the AMP classification model. It consists of a BERT model pretrained on UniRef50 and an MLP model. The input to BERT was the sequences in the AMP and non-AMP data sets, and the output was a vector representation for each sequence. Then, we input the feature representation generated by using BERT into the MLP model. Finally, the classification probability was output after a softmax layer. To visualize the features, we used t-SNE analysis, which gives us a better overview of AMPs (purple points) and non-AMPs (yellow points) that were classified well by the proposed model.

data into a 2D map, thus significantly reducing the perplexity of data. As shown in Figure 3, t-SNE analysis shows that our model can separate data points into two groups of AMPs (purple points) and non-AMPs (yellow points). These results clearly show that our model can address AMP classification efficiently.

Implementation details for the AMP classification model

The development environments and requirements are presented in Table S2. It is essential to advance the predictor's performance by conducting hyperparameter tuning. Some tuning parameters were taken into consideration, i.e. learning rate, batch size and pretrained model. The hyperparameters are optimized according to the maximum area under the curve (AUC). Please refer to Supplementary Table S3 and Figures S2–S4 for more information.

Evaluation metrics for the AMP classification model

In this study, several metrics have been formulated for evaluating the proposed method, which consists of accuracy (Acc), sensitivity (Sens), specificity (Spec), precision (Prec), Mathews correlation coefficient (MCC) and AUC for receiver operating characteristic curve [25, 26]. These metrics are calculated by:

$$\mathrm{Acc} = \frac{\mathrm{TN} + \mathrm{TP}}{\mathrm{TN} + \mathrm{TP} + \mathrm{FN} + \mathrm{FP}} \tag{1}$$

$$\mathrm{Sens} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{2}$$

$$\mathrm{Spec} = \frac{\mathrm{TN}}{\mathrm{TP} + \mathrm{FP}} \tag{3}$$

$$\mathrm{Prec} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{4}$$

$$\mathrm{MCC} = \frac{(\mathrm{TP} \times \mathrm{TN}) - (\mathrm{FP} \times \mathrm{FN})}{\sqrt{(\mathrm{TP} + \mathrm{FP})\,(\mathrm{TP} + \mathrm{FN})\,(\mathrm{TN} + \mathrm{FP})\,(\mathrm{TN} + \mathrm{FN})}} \tag{5}$$

where TP, TN, FP and FN denote the number of true positives, true negatives, false positives and false negatives, respectively.

Peptide synthesis

The peptides mentioned in this study were synthesized by solid-phase peptide synthesis (SPPS) as described previously [45]. Rink amide resin was reacted overnight in a mixed solution of 50% dimethylformamide (DMF)/50% dichloromethane, followed by the addition of 20% piperidine solution for 30 min to remove the Fmoc protecting group from the resin. DMF was used as the solvent for amino acid coupling with the addition of O-(1H-6-chlorobenzotriazole-1-yl)-1,1,3,3-tetramethyluronium hexafluoro-phosphate and N,N-diisopropylethylamine for 1 h at room temperature (20–25 °C). The peptides were cleaved from the resin with a solution of trifluoroacetic acid:triisopropylsilane (Tips):water mixture (90:5:5) for 3 h at room temperature. The peptides were precipitated using ice-cold ether, dissolved in a mixture of H_2O:acetonitrile (1:1) and lyophilized. Crude peptides were purified by RP-HPLC at 214 nm on a Phenomenex C_{18} column, and they were pooled for lyophilization and stored at −20 °C. The sulfhydryl groups of the two cysteines were protected with Acm and reacted in iodine solution at a concentration of 1 mg/mL for 30 min to form disulfide bonds, and the oxidation reaction was finally terminated by adding ascorbic acid. The molecular weights of all peptides were determined using electrospray-mass spectrometry. The purity of the peptides was determined using analytical RP-HPLC, and all purities were >95%.

Nuclear magnetic resonance (NMR) structure determination for A-222

NMR analysis of A-222 was performed in 20 mM sodium phosphate buffer at pH 5.8 with the addition of 5% (v/v) D_2O (90:10). 2D experiments were performed on a Bruker Neo 900 MHz spectrometer at 298 K and consisted of TOCSY, NOESY, ^1H-^{15}N HSQC and ^1H-^{13}C HSQC. TOCSY experiments were also collected at 283–303 K on a Bruker Avance HD III 600 MHz spectrometer. All spectra were assigned with CCPNMR (version 2.4.4). Preliminary structures were calculated using CYANA 3.97 using distance constraints derived from the NOESY spectrum (200 ms mixing time) and backbone dihedral angles generated using the TALOS-N program [46]. The final structural ensemble was generated by the CNS program using torsional angle dynamics, refinement and energy minimization in explicit solvents, and its quality was evaluated using MolProbity [47].

MD simulations

The AMBER16 package and ff14SB force field were used to perform MD simulations to optimize the peptides [48]. The structural stability of the peptides was ensured by MD simulations. The 3D structure of the peptides was generated using AlphaFold2. The peptides were solvated in a 10 Å TIP3P water box, the electrical properties were neutralized using Cl^- and MD simulations were run using the Bash command. The system was optimized by the 2000-step steepest descent and the 3000-step conjugate gradient. After the first energy optimization was completed, the unconstrained optimization was continued, followed by MD simulations. MD simulations include both warming and equilibration processes. First, the systems were gradually heated from 50 to 300 K with the solute restrained to their position using 5 kcal mol^{-1}·Å^{-2} within 100 ps. For the equilibrium process, the solute binding force was gradually reduced from 5 to 0 kcal mol^{-1}·Å^{-2} within 100 ps. Then, 100 ns simulations were conducted at a constant temperature of 300 K and with pressure at 1 atm. The SHAKE algorithm was used for all the hydrogen bonds involved, and a time step of 2 fs was used [49]. After the MD simulation, the MD trajectories were analyzed using VMD (http://www.ks.uiuc.edu/), and root mean square deviation (RMSD) values were calculated. The MD simulations of 46 peptides were performed as described above.

Circular dichroism (CD)

At room temperature in a nitrogen atmosphere, the CD spectrum was measured using a Jasco J-810 spectropolarimeter with wavelengths between 250 and 190 nm, with a 1.0 mm path length cell, 1.0 nm bandwidth and a response time of 2 s, averaging three scans. A-222 was dissolved in a 1:1 mixture of acetonitrile and water at a concentration of 0.3 mg/mL. The spectra are expressed as molar ellipticity. After the measurement, the molar ellipticity [θ] values were calculated, and the secondary structure was analyzed based on the characteristic peaks.

Biological activity assay

The minimum inhibitory concentration (MIC) of AMPs was determined by the broth microdilution method [50]. The tested bacteria included *Bacillus subtilis* 168, *Stenotrophomonas maltophilia* WH 006, *Pseudomonas aeruginosa* SM45, *P. aeruginosa* PAO1, *Bacillus thuringiensis* BNCC 336393, *Staphylococcus aureus* SYZX101, *Escherichia coli* ATCC 8739 and *Lysobacter enzymogenes* YC36. Specifically, serial doubling dilutions of AMPs were performed in 96-well plates (only 50 μL of medium containing diluted AMPs remained in each well). Then, the bacteria cultured to log phase were diluted first to the MacFarlane standard 0.5-fold and then 100-fold. After adding 50 μL of the above bacterial solution to each well, the samples were placed in an incubator for 16–18 h (except *L. enzymogenes* YC36, which was placed at 28 °C, all the others were placed at 37 °C), and the values were recorded at OD_{600}. The experimental group was supplemented with serially diluted AMPs, and medium only (without bacterial solution) and medium containing bacterial solution were used as control groups, with three parallels in each group.

Results and discussion

Estimation of training performance

The learning curves give us better insight into how the learning performance changes over the number of epochs. Meanwhile, they help us diagnose any problems with learning that can lead to an underfit or an overfit model. The curve of training loss and validation loss along with training accuracy and validation accuracy is shown in Figure 4.

As the number of epochs increased, the training accuracy continuously increased, whereas the validation accuracy increased to a maximum after 79 epochs and then began to slightly decrease. The training loss of the model continued to decrease, whereas the validation loss decreased to a minimum and began to increase slowly. This indicates that the model has overfitted after 79 epochs. To avoid overfitting, we saved the model parameters with the highest validation accuracy.

Comparative analysis with state-of-the-art methods

The performance of our model has been compared with various state-of-the-art models on the validation and test sets. According to machine learning algorithms, these predictors can be divided into two categories: (1) predictors constructed based on DNNs, such as Amplify [51] and AMPScannerV2 [22], and (2) predictors that utilize conventional methods to identify AMPs, such as MACREL [52] and AmPEPpy [53].

Comparison using the validation set

The validation set is used for tuning the model's hyperparameters. Meanwhile, it gives us several insights into the performance of the models. As shown in Table 1 and Figure S89A, our proposed model performs better than the other models in most evaluation metrics. In detail, we achieved a sensitivity of 81.86%, accuracy of 85.31%, MCC of 70.79% and AUC of 0.914 in the validation set.

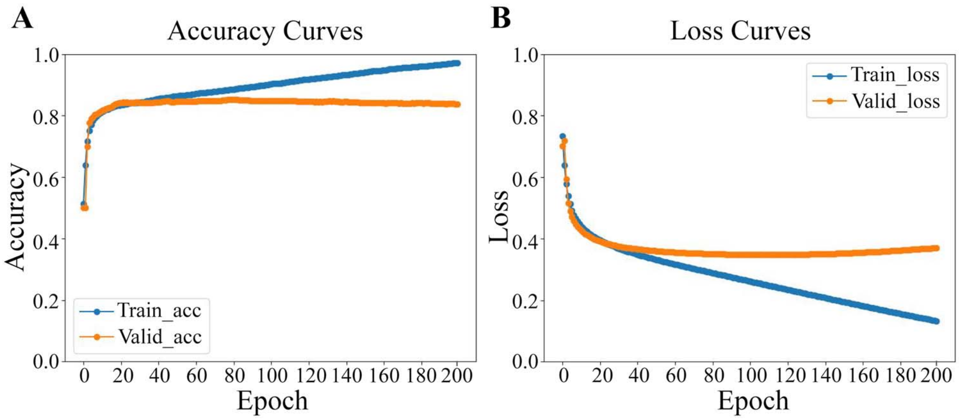

Figure 4. Learning curves of the classification model. (**A**) Accuracy curves. The training accuracy and validation accuracy continue to increase at the beginning, but the validation improves to a maximum after 79 epochs and then decreases slightly. (**B**) Loss curves. The training loss of the model continuously decreases, whereas the validation loss starts to increase slowly after it is reduced to a minimum.

Table 1. Performance of various models using the validation set

Classifier	Acc (%)	Prec (%)	Sens (%)	Spec (%)	MCC (%)	AUC	Algorithm
Ours	**85.31**	87.92	**81.86**	88.75	**70.79**	**0.914**	DNN
Amplify	84.16	**88.97**	77.99	**90.33**	68.84	0.907	DNN
AMPScannerV2	82.83	85.87	78.60	87.06	65.90	0.901	DNN
MACREL	82.35	85.43	77.99	86.70	64.94	0.893	RF
AmPEPpy	79.99	87.01	70.50	89.48	61.09	0.874	RF

The optimal model for each evaluation metrics is given in boldface.

The best performer on Prec and Spec is the Amplify and our model obtains similar performance (with 1.05 and 1.58% lower Prec and Spec values compared with Amplify, respectively). In contrast, our model obtains good performance overall (with 3.87 and 1.95% higher Sens and MCC values compared with Amplify, respectively). The second-best performer on Sens and Acc is the AMPScannerV2, which achieves Sens and Acc values 3.26 and 2.48% lower than our proposed model. It can be noted that all the predictors that utilize conventional methods achieved comparable performances to DNN-based classifier regarding the Prec and Spec metric. This is consistent with recent research that demonstrates that DNN-based models and shallow learning-based models exhibit similar performance [54]. However, according to the MCC metric, our model is 5.58 and 9.7% better than MACREL and AmPEPpy, respectively.

Comparison using independent test set

To test the generalization performance of the models, we utilized an independent test set to compare our proposed model with state-of-the-art methods. As we can see from Table 2 and Figure S89B, the Sens and AUC metrics show that the proposed method outperforms others and achieves the Sens of 95.06% and the AUC of 0.962. It is notable that the second-best predictor in terms of Sens is the conventional method-based classifier MACREL, which is 1.85% lower Sens values compared with our proposed model. Besides, MACREL achieves Acc, Spec and AUC values 3.66, 3.66 and 0.009% lower than our proposed model. The best model in terms of Acc and Spec is Amplify, which achieves Acc and Spec values 5.47 and 5.48% higher than our proposed model. However, according to the Sens metric, our model is 8.02% better than Amplify. The higher sensitivity values can be helpful for us to discover as many AMPs as possible in large-scale screening. Our model, AMPScannerV2 and AmPEPpy perform similarly in terms of Acc and Spec. Specifically, the Acc and Spec values of AMPScannerV2 are 0.74 and 0.76% higher compared with our model, respectively. In terms of Sens and AUC, AMPScannerV2 achieves Sens and AUC values 10.49 and 0.049% lower than our proposed model. In addition, the Acc and Spec values of AmPEPpy are 1.05 and 1.04% higher compared with our model, respectively. In terms of Sens and AUC, AmPEPpy achieves Sens and AUC values 13.58 and 0.049% lower than our proposed model. Besides, the structure of our model is simpler than others. During the training process, we fixed the BERT model weights and only updated the MLP weights using both AMP and non-AMP data sets. Overall, our proposed method is a useful predictor to address AMP classification.

Screening candidate AMPs

Subsequently, our proposed method with the previously existing predictors was applied to AMP recognition (Supplementary Information Table S4). In total, 110 peptides were further screened based on a combination of AlphaFold2 structure prediction and MD simulations. Among these peptides, four peptides contain a pair of disulfide bonds, whereas the rest contain no disulfide bonds. The 3D structures of 110 peptides were predicted by AlphaFold2, and a total of 97 peptides had either α-helix or β-strand components (Supplementary Information Table S4 and Figure S5). The stability of the peptides with complete α-helix or β-strand components was assessed by 50 or 100 ns MD simulations. Most of the peptides exhibited high instability in MD simulations (Figure S6), and their secondary structures were destroyed and even became random coils after 50 ns MD simulations. In addition, we applied the ESMFold [55] model to predict the structure of peptides and perform a comparative analysis regarding the structures predicted by AlphaFold2. In total, 110 peptides were further filtered by our classifier. The SeqGAN-generated peptides were

Table 2. Performance of various models using the independent test set

Classifier	TP	FP	TN	FN	Acc (%)	Sens (%)	Spec (%)	AUC
Ours	154	13 322	98 051	8	88.05	**95.06**	88.04	**0.962**
Amplify	141	7212	104 161	21	**93.52**	87.04	**93.52**	0.955
AMPScannerV2	137	12 473	98 900	25	88.79	84.57	88.80	0.913
MACREL	151	17 402	93 971	11	84.39	93.21	84.38	0.953
AmPEPpy	132	14 474	96 899	30	87.00	81.48	87.00	0.913

The optimal model for each evaluation metrics is given in boldface.

Table 3. Amino acid sequences and physicochemical properties of AMP candidates

Number	Peptide	H-sequence-NH2[a]	MW (Da)	Charge	GRAVY[b]
1	A-5	FLGPLISLLLCGQLTCKL	1929.46	+1	1.533
2	A-19	KLQIMKKFRSIANKTMNT	2151.67	+5	−0.644
3	A-72	REELQRVVKRFVVVQQLS	2212.64	+2	−0.239
4	A-90	FLQGIKACRKFSCRRAVS	2067.51	+5	−0.006
5	A-222	DTFGRCRRWWAALGACRR	2178.54	+4	−0.683
6	A-252	FLPLICRSQCIRSLGTGP	1958.38	+2	0.522

[a]The two cysteines of the peptides shown in the table form a disulfide bond. [b]The GRAVY score (grand average of hydropathicity) is a method for calculating the hydrophobicity/solubility of a peptide A negative GRAVY score indicates hydrophilicity, whereas a positive GRAVY score indicates hydrophobicity.

predicted by AlphaFold2 as well as ESMFold, as shown in Table S5. Among the 110 peptides identified as AMPs by our classifier, a total of 97 peptides were predicted to have either α-helix or β-strand components by AlphaFold2, whereas 83 peptides were predicted to have secondary structure by ESMFold. Then, we performed a comparative analysis regarding the structures predicted by AlphaFold2 and ESMFold. It was found that 82 of them could get the same results (had either α-helix or β-strand components) using both. This also indicated that the functionality of the AMPs is highly correlated with the presence of the secondary structures [31, 55]. Please refer to Supplementary Information Tables S4 and S5 for more information. We selected six peptides that could fully or partially maintain their secondary structure and showed a relatively small RMSD as candidate AMPs for wet laboratory experimental validation (Figure 5A and B).

Wet laboratory

Synthesis of AMP candidates

AMP candidates were synthesized by SPPS and purified by RP-HPLC. The mass and purity were validated using electrospray-mass and analytical RP-HPLC, respectively (Figures S8–S19). We studied the physicochemical properties of all AMP candidates (Table 3). The molecular weight of the peptides ranged from 1929.46 Da (1) to 2212.64 Da (3). The charge of the AMP candidates varied from +1 (1) to +5 (2, 4). Peptides 2– 5 were predicted to be hydrophilic based upon their GRAVY score [56], whereas the remaining peptides were predicted to be hydrophobic.

Antimicrobial activity

These peptides were tested for antimicrobial activity, as determined by the MIC against gram-positive strains (*B. subtilis* 168, *S. aureus* SYZX101 and *B. thuringiensis* BNCC 336393) and gram-negative strains (*E. coli* ATCC 8739, *L. enzymogenes* YC36, *S. maltophilia* WH 006, *P. aeruginosa* PAO1 and *P. aeruginosa* SM45). Among these candidate AMPs, we found that A-222 (DTFGRCRRWWAALGACRR-NH_2) showed highly effective broad-spectrum antimicrobial activity, with MIC values of 16 μg/mL against *B. subtilis* 168 and *L. enzymogenes* YC36, as well as MIC values of 32 and 64 μg/mL against *S. maltophilia* WH 006 and *P. aeruginosa* PAO1, respectively (Table 4).

Structural analyses

NMR is an important tool for analyzing the 3D structures of peptides and proteins. In particular, previous work has highlighted the significance of α-helices and amphiphilic structures for the activity of AMPs [57, 58]. In this study, the solution structure of the most active peptide, A-222, was determined. AlphaFold2 predicted its structure as a several-turn α-helix with a flexible N-terminal tail (Figure 6A), whereas the 10 frames extracted from the 100 ns MD simulation demonstrated the stability of this predicted structure (Figure 6B). In addition, ESMFold predicted its structure, as shown in Figure S7. Then, we performed NMR structural analysis of A-222 (Figure 6C and D). Secondary alpha proton shift analysis revealed a negative shift of all residues relative to random coil chemical shifts,. whereas structure calculations confirmed the presence of a small alpha helix across residues 10–13. Only a single hydrogen bond was predicted, based upon preliminary structures and variable temperature amide coefficients, between the amide proton and carbonyl oxygen of residues 14 and 10, respectively, which provides further evidence that only a short helix is present in the NMR structure. Details of the restraints and structural statistics are presented in Table S6. The discrepancy between the NMR structure and the AlphaFold2 predicted structure might originate from the different solution environments. The NMR structure was determined in sodium phosphate buffer at pH 5.8, which might affect the folding of the peptide. In addition, we also performed CD spectroscopy [59]. The characteristic peaks of α-helices are negative at 222 and 208 nm and positive near 190 nm, which also confirmed the presence of the A-222 α-helix (Figure 6E). Thus, A-222 maintains its secondary structure, and this stabilized structure might be relevant for antibacterial activity.

Design of A-222 analogs and biological activity assay

In previous MIC tests of A-222, it showed broad-spectrum antibacterial activity against both gram-positive and gram-negative bacteria, particularly against *L. enzymogenes* YC36, *B. subtilis* 168, *S. maltophilia* WH 006 and *P. aeruginosa* PAO1. We focused on these four bacterial strains for bioactivity testing to assess the effect of modification on A-222. Alanine scanning is a common method to study SARs and is often used to determine the effect of each

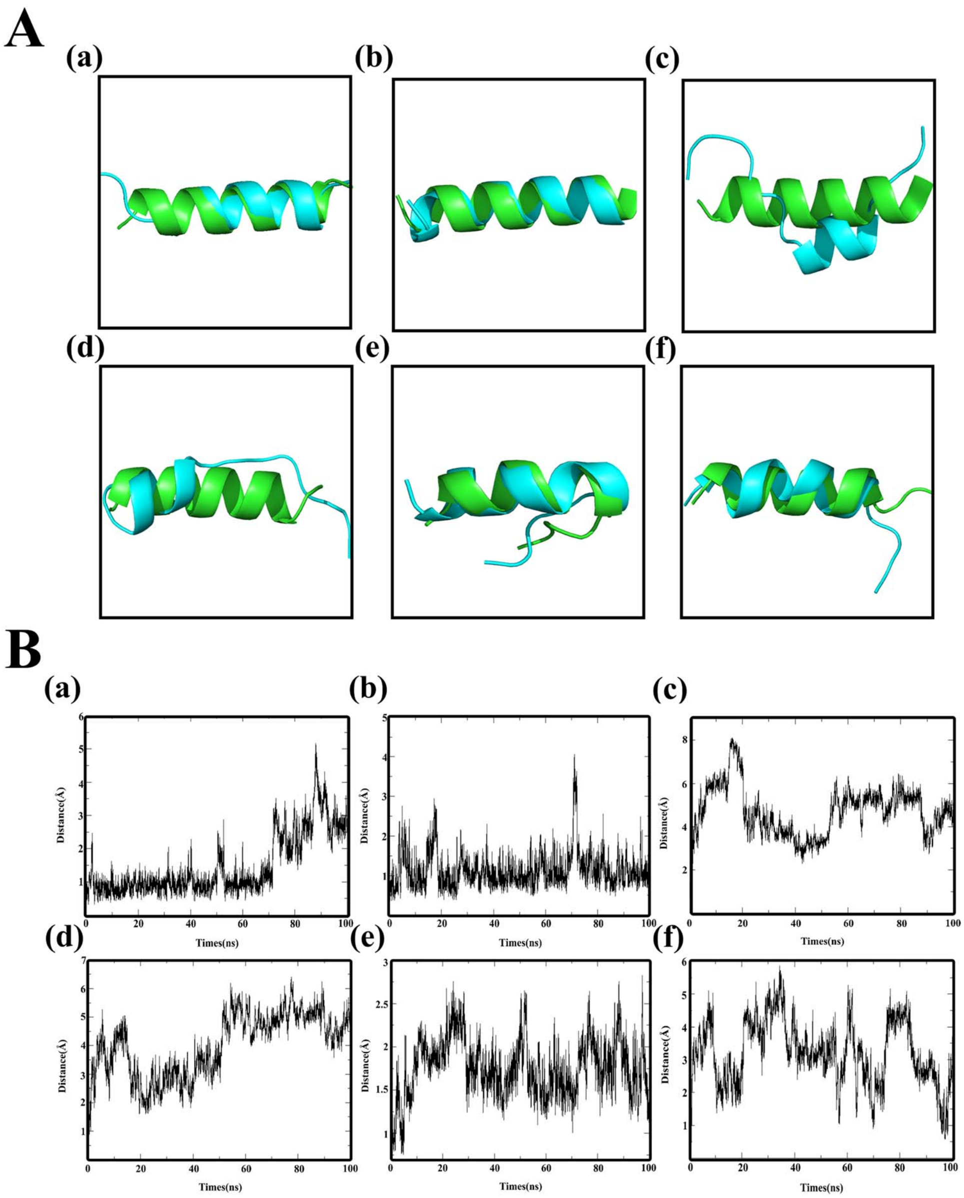

Figure 5. MD simulation of candidate AMPs. (**A**) The structure predicted by AlphaFold2 (in green) was aligned with the structure from MD simulations (in cyan) performed at 100 ns. (**B**) RMSD of AMP candidates in the 100 ns MD simulation. (a, b, c, d, e, f): A-19, A-72, A-5, A-90, A-222, A-252.

Table 4. MICs of candidate AMPs against bacterial pathogens[a]

MIC (μg/mL)[a]								
Compound	**Gram-positive bacteria**			**Gram-negative bacteria**				
	B. subtilis	***S. aureus***	***B. thuringiensis***	***E. coli***	***S. maltophilia***	***L. enzymogenes***	***P. aeruginosa***	
	168	**SYZX101**	**BNCC 336393**	**ATCC 8739**	**WH 006**	**YC36**	**SM45**	**PAO1**
A-5	>256	>256	>256	>256	>256	>256	>256	>256
A-19	>256	>256	>256	64	128	64	256	128
A-72	>256	>256	>256	>256	>256	>256	>256	>256
A-90	>256	>256	>256	256	>256	>256	>256	>256
A-222	16	128	256	64	32	16	128	64
A-252	>256	>256	>256	>256	>256	>256	>256	>256

[a]The MIC was tested in triplicate.

amino acid residue on physicochemical properties and biological activity [60] (Table 5). In an alanine scan of A-222 (Figures S20–S45), it was found that the D1A mutation contributed to a 2-fold increased activity against *L. enzymogenes* YC36 and *S. maltophilia* WH 006. Furthermore, the MIC value was significantly reduced from 64 to 16 μg/mL against *P. aeruginosa* PAO1. Removal

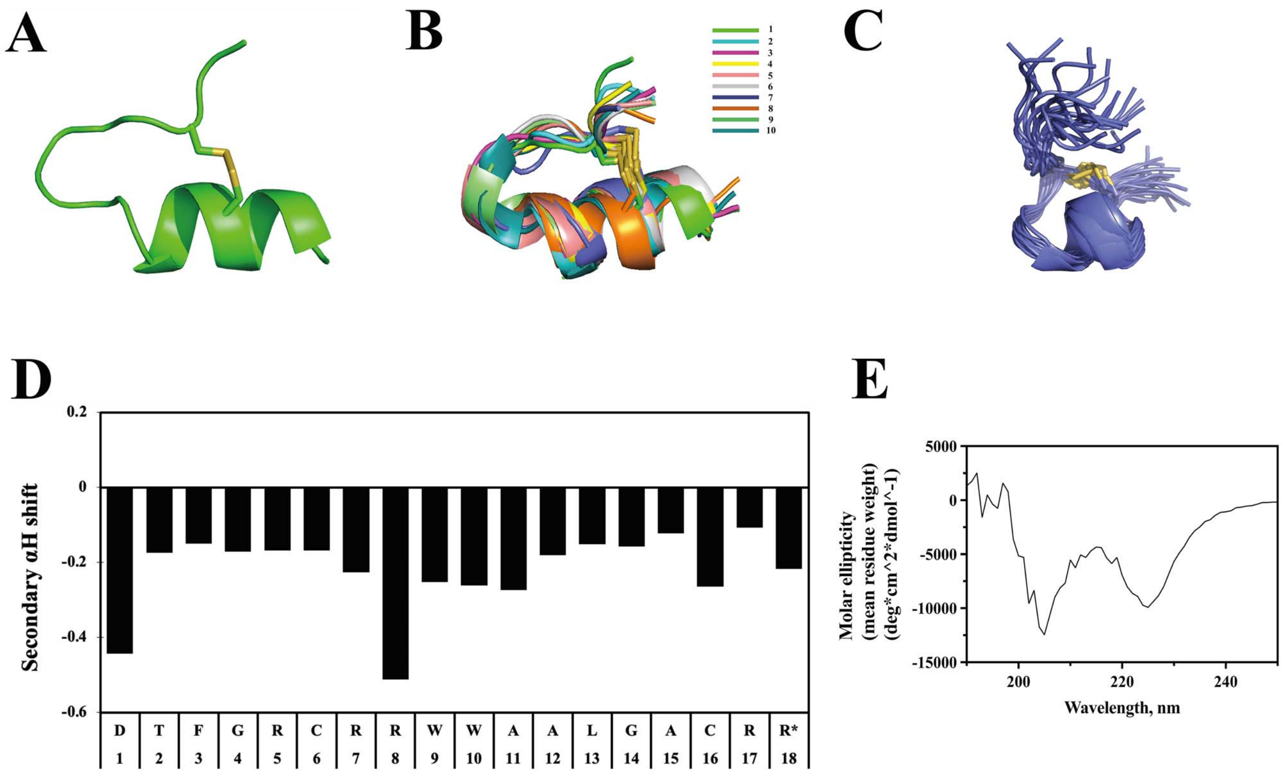

Figure 6. Structural analysis of A-222. (**A**) Structure of A-222 based on AlphaFold2 prediction. (**B**) Ten frames extracted from 100 ns MD simulations of the AlphaFold2 predicted structure. (**C**) NMR solution structure as an overlay of the 20 lowest-energy states. The disulfide bond is shown in yellow. (**D**) Secondary proton chemical shift analysis. The asterisk denotes a C-terminal amide. (**E**) CD spectrum.

of the negatively charged sidechain of aspartate likely explains the activity change [61]. However, the A-222[D1A] mutation did not induce significant changes in gram-positive bacteria, such as *B. subtilis* 168 (Table 6).

To further explore the effect of an increased positive charge, we next performed arginine scanning of A-222 (Figures S46–S67 and Table 7). Positively charged residues play a key role in biological activity, and adding a positive charge facilitates the binding of AMPs to the bacterial membrane, thus improving antimicrobial activity [62]. The activity of our mutants against *S. maltophilia* WH 006 was not increased as expected during arginine scanning. However, we found that when amino acids at positions 1, 3, 11 and 12 were replaced with arginine, the activity was increased 4–8-fold against *P. aeruginosa* PAO1 (Table 8). *P. aeruginosa* is also a clinically common pathogenic bacterium, and infections caused by *P. aeruginosa* involve a high mortality rate and are difficult to treat [63]. The change indicates that the increase in activity can indeed be achieved by changing the amino acids at certain positions through increasing the number of positive charges. Therefore, adding positively charged residues in AMPs is an effective strategy to improve the activity against *P. aeruginosa* PAO1.

In the third round of modification, we continued with the strategy of increasing the positive charges. We chose positions that were not adjacent sequentially or spatially and designed double mutants of A-222 (Table 9 and Figures S68–77 and S88). On this basis, we then designed triple and quadruple mutants of A-222 by combining the positions that could potentially increase the potency of the peptide (Table 10 and Figures S78–S87). For *P. aeruginosa* PAO1, the results showed that when the activity reached a certain value, adding further positive charges did not increase the activity. Conversely, the activity was decreased to some extent, which might originate from the excessive positive charges that change the amphiphilicity of the peptides. Nevertheless, the activity of the modified peptide was higher than that of A-222 before modification. Amphiphilicity is among the important factors that determine whether a peptide can function as an antibacterial agent or not [64]. Interestingly, this strategy of increasing the number of positive charges can significantly increase the activity against *S. maltophilia* WH 006. *S. maltophilia* that exhibits significant intrinsic antibiotic resistance can cause chronic pulmonary infections in several immunocompromised patients [65]. These AMPs that showed promising activity against *S. maltophilia* WH 006 had charge numbers of +6 or +7 and GRAVY values in the approximate range of −1 to −1.3. In brief, AMPs with such physical characteristics are beneficial in inhibiting the growth of *S. maltophilia* WH 006. In this modification, we also tested another common gram-positive bacterium, *B. subtilis* 168. However, no significant change in activity was found (Table 10).

Conclusions

Here, we designed and identified AMPs by utilizing SeqGAN, BERT and MLP. Previous experimental studies demonstrated that the function of AMPs is highly correlated with the stability of the secondary structures, such as helix or beta-strand components; thus, we used the availability of the secondary structure and its stability as an important basis to further select the generated peptides [32, 33]. Eventually, six candidate AMPs were successfully designed after a combination of MD and AlphaFold2. Among them, A-222 showed high potency against a variety of gram-negative and gram-positive bacteria. The presence of an alpha

Table 5. Alanine mutants of A-222

Number	Peptide	H-sequence-NH2[a]	MW (Da)	Charge	GRAVY
1	A-222	DTFGRCRRWWAALGACRR	2178.54	+4	–0.683
2	A-222[D1A]	**A**TFGRCRRWWAALGACRR	2134.53	+5	–0.389
3	A-222[T2A]	D**A**FGRCRRWWAALGACRR	2148.51	+4	–0.544
4	A-222[F3A]	DT**A**GRCRRWWAALGACRR	2102.44	+4	–0.739
5	A-222[G4A]	DTF**A**RCRRWWAALGACRR	2192.56	+4	–0.561
6	A-222[R5A]	DTFG**A**CRRWWAALGACRR	2093.43	+3	–0.333
7	A-222[R7A]	DTFGRC**A**RWWAALGACRR	2093.43	+3	–0.333
8	A-222[R8A]	DTFGRCR**A**WWAALGACRR	2093.43	+3	–0.333
9	A-222[W9A]	DTFGRCRR**A**WAALGACRR	2063.40	+4	–0.533
10	A-222[W10A]	DTFGRCRRW**A**AALGACRR	2063.40	+4	–0.533
11	A-222[L13A]	DTFGRCRRWWAA**A**GACRR	2136.46	+4	–0.794
12	A-222[G14A]	DTFGRCRRWWAAL**A**ACRR	2192.56	+4	–0.561
13	A-222[R17A]	DTFGRCRRWWAALGAC**A**R	2093.43	+3	–0.333
14	A-222[R18A]	DTFGRCRRWWAALGACR**A**	2093.43	+3	–0.333

[a]The two cysteines of the peptides shown in the table form a disulfide bond. Bold-underline letters represent amino acids that were replaced from A-222.

Table 6. Antimicrobial activity of A-222 and its alanine scanning mutants[a]

MIC (μg/mL)[a]				
Compound	**Gram-positive bacteria**	**Gram-negative bacteria**		
	B. subtilis	***S. maltophilia***	***L. enzymogenes***	***P. aeruginosa***
	168	**WH 006**	**YC36**	**PAO1**
A-222[D1A]	16	16	8	16
A-222[T2A]	16	>16	16	>16
A-222[F3A]	16	>16	>16	>16
A-222[G4A]	16	>16	>16	>16
A-222[R5A]	>16	>16	>16	>16
A-222[R7A]	>16	>16	>16	>16
A-222[R8A]	>16	>16	>16	>16
A-222[W9A]	>16	>16	>16	>16
A-222[W10A]	>16	>16	>16	>16
A-222[L13A]	>16	>16	>16	>16
A-222[G14A]	16	>16	>16	>16
A-222[R17A]	>16	>16	>16	>16
A-222[R18A]	>16	>16	>16	>16

[a]The MIC was tested in triplicate.

Table 7. Amino acid sequence of A-222 arginine scanning

Number	Peptide	H-sequence-NH2[a]	MW (Da)	Charge	GRAVY
1	A-222	DTFGRCRRWWAALGACRR	2178.54	+4	–0.683
2	A-222[D1R]	**R**TFGRCRRWWAALGACRR	2219.64	+6	–0.739
3	A-222[T2R]	D**R**FGRCRRWWAALGACRR	2233.62	+5	–0.894
4	A-222[F3R]	DT**R**GRCRRWWAALGACRR	2187.55	+5	–1.089
5	A-222[G4R]	DTF**R**RCRRWWAALGACRR	2260.67	+5	–0.911
6	A-222[W9R]	DTFGRCRR**R**WAALGACRR	2131.51	+5	–0.883
7	A-222[W10R]	DTFGRCRRW**R**AALGACRR	2131.51	+5	–0.883
8	A-222[A11R]	DTFGRCRRWW**R**ALGACRR	2263.65	+5	–1.033
9	A-222[A12R]	DTFGRCRRWWA**R**LGACRR	2263.65	+5	–1.033
10	A-222[L13R]	DTFGRCRRWWAA**R**GACRR	2221.57	+5	–1.144
11	A-222[G14R]	DTFGRCRRWWAAL**R**ACRR	2277.67	+5	–0.911
12	A-222[A15R]	DTFGRCRRWWAALG**R**CRR	2263.65	+5	–1.033

[a]The two cysteines of the peptides shown in the table form a disulfide bond. Bold-underline letters represent amino acids that were replaced from A-222.

helix of A-222 was demonstrated by NMR structure studies, CD conformation analysis and AlphaFold2 prediction. However, it is noted that the structure of A-222 predicted based on AlphaFold2 is substantially different from the structure determined using NMR despite the presence of the alpha helix in both structures. The alpha helix is much longer in the AlphaFold2 predicted structure than in the NMR structure. The solvent environment might affect the folding of the peptide. Subsequently, we designed and synthesized new analogs of A-222 and evaluated the effect on antimicrobial activity by performing alanine scanning as well as

Table 8. Antimicrobial activity of A-222 and its arginine scanning analogs[a]

MIC (μg/mL)[a]				
Compound	Gram-positive bacteria	Gram-negative bacteria		
	B. subtilis	*S. maltophilia*	*L. enzymogenes*	*P. aeruginosa*
	168	WH 006	YC36	PAO1
A-222[D1R]	32	32	8	16
A-222[T2R]	>32	>32	8	32
A-222[F3R]	>32	>32	16	8
A-222[G4R]	>32	>32	8	32
A-222[W9R]	>32	>32	>32	>32
A-222[W10R]	>32	>32	>32	>32
A-222[A11R]	>32	>32	16	16
A-222[A12R]	>32	>32	8	8
A-222[L13R]	>32	>32	>32	>32
A-222[G14R]	>32	>32	16	32
A-222[A15R]	>32	>32	16	32

[a] The MIC was tested in triplicate.

Table 9. Amino acid sequences of A-222 analogs[a]

Number	**Peptide**	**H-sequence-NH2[a]**	**MW (Da)**	**Charge**	**GRAVY**
1	A-222	DTFGRCRRWWAALGACRR	2178.54	+4	−0.683
2	A-222[D1R,A11R]	**R**TFGRCRRWW**R**ALGACRR	2305.73	+7	−1.089
3	A-222[D1R,G14R]	**R**TFGRCRRWWAAL**R**ACRR	2318.78	+7	−0.967
4	A-222[D1R,A15R]	**R**TFGRCRRWWAALG**R**CRR	2305.73	+7	−1.089
5	A-222[A11R,G14R]	DTFGRCRRWW**R**AL**R**ACRR	2362.78	+6	−1.261
6	A-222[A11R,A15R]	DTFGRCRRWW**R**ALG**R**CRR	2348.76	+6	−1.383
7	A-222[D1R,A11R,A15R]	**R**TFGRCRRWW**R**ALG**R**CRR	2389.86	+8	−1.439
8	A-222[D1R,A11R,G14R]	**R**TFGRCRRWW**R**AL**R**ACRR	2403.89	+8	−1.317
9	A-222[D1R,G14R,A15R]	**R**TFGRCRRWWAAL**RR**CRR	2403.89	+8	−1.317
10	A-222[A11R,G14R,A15R]	DTFGRCRRWW**R**AL**RR**CRR	2447.89	+7	−1.611
11	A-222[4R]	**R**TFGRCRRWW**R**AL**RR**CRR	2489.00	+9	−1.667

[a] The two cysteines of the peptides shown in the table form a disulfide bond. Bold-underline letters represent amino acids that were replaced from A-222.

Table 10. Antimicrobial activity of A-222 analogs[a]

MIC (μg/mL)[a]				
Compound	Gram-positive bacteria	Gram-negative bacteria		
	B. subtilis	*S. maltophilia*	*L. enzymogenes*	*P. aeruginosa*
	168	WH 006	YC36	PAO1
A-222[D1R,A11R]	>16	8	16	8
A-222[D1R,G14R]	>16	8	16	8
A-222[D1R,A15R]	>16	8	16	16
A-222[A11R,G14R]	>16	8	8	8
A-222[A11R,A15R]	>16	8	16	16
A-222[D1R,A11R,A15R]	>16	8	16	16
A-222[D1R,A11R,G14R]	16	>16	8	16
A-222[D1R,G14R,A15R]	16	16	16	16
A-222[A11R,G14R,A15R]	>16	16	16	16
A-222[4R]	16	>16	16	16

[a] The MIC was tested in triplicate.

increasing the number of positively charged residues. Subsequent modification of A-222 revealed a 4–8-fold increase in activity against *S. maltophilia* WH 006 and *P. aeruginosa* PAO1, and the MIC for several analogs reached 8 μg/mL, which is comparable with that of naturally occurring AMPs. Our work confirms the potential of deep learning and MD simulations in accelerating AMP discovery.

Key Points

- Discovering and designing AMPs using traditional methods is a time-consuming and costly process. Deep learning has been applied to the *de novo* design of AMPs and address AMP classification with high efficiency.

- In this study, we established a flowchart for the efficient generation of AMPs by combining deep learning and MD simulations with wet laboratory experimental studies. Candidate AMPs were successfully selected after a combination of MD and Alphafold2.
- In a wet laboratory, we found that A-222 showed high potency against a variety of gram-negative and gram-positive bacteria.
- We then performed a structure–activity relationship study to design a new series of peptide analogs and found that the activities of these analogs could be increased by 4–8-fold against *Stenotrophomonas maltophilia* WH 006 and *Pseudomonas aeruginosa* PAO1.

Supplementary data

Supplementary data are available online at https://academic.oup.com/bib.

Author contributions

R.Y., S.C. and Y.W. designed and managed the project. C.G. designed the sequence generation with help from S.C. and analyzed generated sequences. Q.C. performed and analyzed the molecular dynamics simulations. Q.C., X.J.W., Z. Z. and Y.M. performed and analyzed the wet-laboratory experiments. NMR was performed by P.J.H. and X.Y.J. All of the authors contributed to writing and revision of the manuscript.

Code availability

The code is available at GitHub (https://github.com/gc-js/Antimicrobial-peptide-generation). The users can submit their own sequences and get prediction results at HuggingFace Spaces (https://huggingface.co/spaces/oucgc1996/Antimicrobial-peptide-generation).

Acknowledgements

This work was supported by the National Natural Science Foundation of China (NSFC) (Nos 82122064, 32171267, 42176108), a grant from the National Science and Technology Major Project for Significant New Drugs Development (No. 2018ZX09735004) and Qingdao National Laboratory for Marine Science and Technology (No. 2022QNLM030003). We would like to thank Primary Biotech (www.pumeirui.com) for the support of molecular modeling. P.J.H. and D.J.C. acknowledge funding by the Australian Research Council Centre of Excellence for Innovations in Peptide and Protein Science (CE200100012). D.J.C. is supported by a National Health and Medical Research Council (Australia) Investigator Grant (2009564).

References

1. Ventola CL. The antibiotic resistance crisis: part 1: causes and threats. *P t* 2015;**40**:277–83.
2. Ma Y, Guo Z, Xia B, *et al.* Identification of antimicrobial peptides from the human gut microbiome using deep learning. *Nat Biotechnol* 2022;**40**:921–31.
3. Mishra B, Reiling S, Zarena D, *et al.* Host defense antimicrobial peptides as antibiotics: design and application strategies. *Curr Opin Chem Biol* 2017;**38**:87–96.
4. Roncevic T, Puizina J, Tossi A. Antimicrobial peptides as anti-infective agents in pre-post-antibiotic era? *Int J Mol Sci* 2019; **20**:5713.
5. Yu G, Baeder DY, Regoes RR, *et al.* Predicting drug resistance evolution: insights from antimicrobial peptides and antibiotics. *Proc Biol Sci* 2018;**285**:20172687.
6. Zhang LJ, Gallo RL. Antimicrobial peptides. *Curr Biol* 2016;**26**: R14–9.
7. Lazzaro BP, Zasloff M, Rolff J. Antimicrobial peptides: application informed by evolution. *Science* 2020;**368**:487.
8. Usmani SS, Bedi G, Samuel JS, *et al.* THPdb: database of FDA-approved peptide and protein therapeutics. *PLoS One* 2017; **12**:e0181748.
9. Zhang L, Wang CC, Chen X. Predicting drug-target binding affinity through molecule representation block based on multi-head attention and skip connection. *Brief Bioinform* 2022;**23**:1–12.
10. Chen X, Sun L-G, Zhao Y. NCMCMDA: miRNA–disease association prediction through neighborhood constraint matrix completion. *Brief Bioinform* 2020;**22**:485–96.
11. Chen X, Li T-H, Zhao Y, *et al.* Deep-belief network for predicting potential miRNA-disease associations. *Brief Bioinform* 2020;**22**: 1–10.
12. Lipinski CF, Maltarollo VG, Oliveira PR, *et al.* Advances and perspectives in applying deep learning for drug design and discovery. *Front Robot AI* 2019;**6**:108.
13. Gao Y, Fang H, Fang L, *et al.* The modification and design of antimicrobial peptide. *Curr Pharm Design* 2018;**24**:904–10.
14. Das P, Sercu T, Wadhawan K, *et al.* Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nat Biomed Eng* 2021;**5**:613–23.
15. Surana S, Arora P, Singh D, *et al.* PandoraGAN: generating antiviral peptides using generative adversarial network. *bioRxiv* 2021; 2021.02.15.431193. https://doi.org/10.1101/2021.02.15.431193.
16. Tucs A, Tran DP, Yumoto A, *et al.* Generating ampicillin-level antimicrobial peptides with activity-aware generative adversarial networks. *ACS Omega* 2020;**5**:22847–51.
17. Dean SN, Alvarez JAE, Zabetakis D, *et al.* PepVAE: variational autoencoder framework for antimicrobial peptide generation and activity prediction. *Front Microbiol* 2021;**12**:725727.
18. Liang Y, Zhang X, Yuan Y, *et al.* Role and modulation of the secondary structure of antimicrobial peptides to improve selectivity. *Biomater Sci* 2020;**8**:6858–66.
19. Pinacho-Castellanos SA, García-Jacas CR, Gilson MK, *et al.* Alignment-free antimicrobial peptide predictors: improving performance by a thorough analysis of the largest available data set. *J Chem Inf Model* 2021;**61**:3141–57.
20. Singh V, Shrivastava S, Singh SK, *et al.* StaBle-ABPpred: a stacked ensemble predictor based on biLSTM and attention mechanism for accelerated discovery of antibacterial peptides. *Brief Bioinform* 2022;**23**:1–17.
21. Jan A, Hayat M, Wedyan M, *et al.* Target-AMP: computational prediction of antimicrobial peptides by coupling sequential information with evolutionary profile. *Comput Biol Med* 2022;**151**:106311.
22. Veltri D, Kamath U, Shehu A. Deep learning improves antimicrobial peptide recognition. *Bioinformatics* 2018;**34**:2740–7.
23. Yan K, Lv H, Guo Y, *et al.* sAMPpred-GAT: prediction of antimicrobial peptide by graph attention network and predicted peptide structure. *Bioinformatics* 2022;**39**:btac715.
24. Hussain W. sAMP-PFPDeep: improving accuracy of short antimicrobial peptides prediction using three different sequence encodings and deep neural networks. *Brief Bioinform* 2022;**23**: 1–12.

25. Kha Q-H, Ho Q-T, Le NQK. Identifying SNARE proteins using an alignment-free method based on multiscan convolutional neural network and PSSM profiles. *J Chem Inf Model* 2022;**62**: 4820–6.
26. Le NQK, Ho Q-T, Nguyen V-N, *et al.* BERT-promoter: an improved sequence-based predictor of DNA promoter using BERT pre-trained model and SHAP feature selection. *Comput Biol Chem* 2022;**99**:107732.
27. Naseer S, Hussain W, Khan YD, *et al.* Optimization of serine phosphorylation prediction in proteins by comparing human engineered features and deep representations. *Anal Biochem* 2021;**615**:114069.
28. Lin Z, Akin H, Rao R, *et al.* Language models of protein sequences at the scale of evolution enable accurate structure prediction. *bioRxiv* 2022; 2022.07.20.500902. https://doi.org/10.1101/2022.07.20.500902.
29. Sanches K, Wai DCC, Norton RS. Conformational dynamics in peptide toxins: implications for receptor interactions and molecular design. *Toxicon* 2021;**201**:127–40.
30. Cramer P. AlphaFold2 and the future of structural biology. *Nat Struct Mol Biol* 2021;**28**:704–5.
31. Jumper J, Evans R, Pritzel A, *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* 2021;**596**: 583–9.
32. Yokoo H, Hirano M, Ohoka N, *et al.* Structure-activity relationship study of amphipathic antimicrobial peptides using helix-destabilizing sarcosine. *J Pept Sci* 2021;**27**: e3360.
33. Hirano M, Saito C, Goto C, *et al.* Rational design of helix-stabilized antimicrobial peptide foldamers containing alpha,alpha-disubstituted amino acids or side-chain stapling. *ChemPlusChem* 2020;**85**:2731–6.
34. Lee HT, Lee CC, Yang JR, *et al.* A large-scale structural classification of antimicrobial peptides. *Biomed Res Int* 2015;**2015**:475062.
35. Gawde U, Chakraborty S, Waghu FH, *et al.* CAMPR4: a database of natural and synthetic antimicrobial peptides. *Nucleic Acids Res* 2022;**51**:D377–83.
36. Aguilera-Mendoza L, Marrero-Ponce Y, Garcia-Jacas CR, *et al.* Automatic construction of molecular similarity networks for visual graph mining in chemical space of bioactive peptides: an unsupervised learning approach. *Sci Rep* 2020;**10**:18074.
37. Sidorczuk K, Gagat P, Pietluch F, *et al.* Benchmarks in antimicrobial peptide prediction are biased due to the selection of negative data. *Brief Bioinform* 2022;**23**:1–12.
38. Burdukiewicz M, Sidorczuk K, Rafacz D, *et al.* Proteomic screening for prediction and design of antimicrobial peptides with AmpGram. *Int J Mol Sci* 2020;**21**:4310.
39. Fu L, Niu B, Zhu Z, *et al.* CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 2012;**28**:3150–2.
40. Wang G, Li X, Wang Z. APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res* 2016;**44**:D1087–93.
41. Waghu FH, Barai RS, Gurung P, *et al.* CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. *Nucleic Acids Res* 2016;**44**:D1094–7.
42. Yu LT, Zhang WN, Wang J *et al.* SeqGAN: sequence generative adversarial nets with policy gradient. In: *31st Association for the Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence*, 2017, p. 2852–8. San Francisco, CA.
43. García-Jacas CR, García-González LA, Martinez-Rios F, *et al.* Handcrafted versus non-handcrafted (self-supervised) features for the classification of antimicrobial peptides: complementary or redundant? *Brief Bioinform* 2022;**23**:1–16.
44. Laurens VDM, Hinton G. Visualizing data using t-SNE. *J Mach Learn Res* 2008;**9**:2579–605.
45. Yu R, Wang J, So LY, *et al.* Enhanced activity against multidrug-resistant bacteria through coapplication of an analogue of tachyplesin I and an inhibitor of the QseC/B SIGNALING pathway. *J Med Chem* 2020;**63**:3475–84.
46. Shen Y, Bax A. Protein backbone and sidechain torsion angles predicted from NMR chemical shifts using artificial neural networks. *J Biomol NMR* 2013;**56**:227–41.
47. Chen VB, Arendall WB, 3rd, Headd JJ, *et al.* MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr D Biol Crystallogr* 2010;**66**:12–21.
48. Maier JA, Martinez C, Kasavajhala K, *et al.* ff14SB: improving the accuracy of protein side chain and backbone parameters from ff99SB. *J Chem Theory Comput* 2015;**11**:3696–713.
49. Miyamoto S, Kollman PA. Settle: an analytical version of the SHAKE and RATTLE algorithm for rigid water models. *J Comput Chem* 2010;**13**:952–62.
50. Wiegand I, Hilpert K, Hancock RE. Agar and broth dilution methods to determine the minimal inhibitory concentration (MIC) of antimicrobial substances. *Nat Protoc* 2008;**3**:163–75.
51. Li C, Sutherland D, Hammond SA, *et al.* AMPlify: attentive deep learning model for discovery of novel antimicrobial peptides effective against WHO priority pathogens. *BMC Genomics* 2022;**23**:77.
52. Santos-Júnior C, Pan S, Zhao XM, *et al.* Macrel: antimicrobial peptide screening in genomes and metagenomes. *PeerJ* 2020;**8**:e10555.
53. Lawrence TJ, Carper DL, Spangler MK, *et al.* amPEPpy 1.0: a portable and accurate antimicrobial peptide prediction tool. *Bioinformatics* 2020;**37**:2058–60.
54. García-Jacas CR, Pinacho-Castellanos SA, García-González LA, *et al.* Do deep learning models make a difference in the identification of antimicrobial peptides? *Brief Bioinform* 2022;**23**:1–16.
55. Rives A, Meier J, Sercu T, *et al.* Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc Natl Acad Sci U S A* 2021;**118**:e2016239118.
56. Artimo P, Jonnalagedda M, Arnold K, *et al.* ExPASy: SIB bioinformatics resource portal. *Nucleic Acids Res* 2012;**40**:W597–603.
57. Vincenzi M, Mercurio AF, Leone M. NMR spectroscopy in the conformational analysis of peptides: an overview. *Curr Med Chem* 2021;**28**:2729–82.
58. Benetti S, Timmons PB, Hewage CM. NMR model structure of the antimicrobial peptide maximin 3. *Eur Biophys J* 2019;**48**:203–12.
59. Greenfield NJ. Using circular dichroism spectra to estimate protein secondary structure. *Nat Protoc* 2006;**1**:2876–90.
60. Migon D, Jaskiewicz M, Neubauer D, *et al.* Alanine scanning studies of the antimicrobial peptide Aurein 1.2, probiotics Antimicrob. *Proteins* 2019;**11**:1042–54.
61. Cantini F, Luzi C, Bouchemal N, *et al.* Effect of positive charges in the structural interaction of crabrolin isoforms with lipopolysaccharide. *J Pept Sci* 2020;**26**:e3271.
62. Aschi M, Perini N, Bouchernal N, *et al.* Structural characterization and biological activity of crabrolin peptide isoforms with different positive charge. *BBA-Biomembranes* 2020;**1862**: 183055.
63. Kerr KG, Snelling AM. *Pseudomonas aeruginosa*: a formidable and ever-present adversary. *J Hosp Infect* 2009;**73**:338–44.
64. Torres MDT, Sothiselvam S, Lu TK, *et al.* Peptide design principles for antimicrobial applications. *J Mol Biol* 2019;**431**:3547–67.
65. Trifonova A, Strateva T. *Stenotrophomonas maltophilia* - a low-grade pathogen with numerous virulence factors. *Infect Dis* 2019;**51**:168–78.