OXFORD

Briefings in Bioinformatics, 2025, **26(1)**, bbae715
https://doi.org/10.1093/bib/bbae715
Problem Solving Protocol

Directed evolution of antimicrobial peptides using multi-objective zeroth-order optimization

Xianliang Liu[1], Jiawei Luo[1], Xinyan Wang[2], Yang Zhang[3], Junjie Chen[1],*

[1]School of Computer Science and Technology, Harbin Institute of Technology, HIT Campus, Shenzhen University Town, Nanshan District, Shenzhen 518055, Guangdong, China
[2]Core Research Facility, Southern University of Science and Technology, No. 1088 Xueyuan Road, Nanshan District, Shenzhen 518055, Guangdong, China
[3]School of Science, Harbin Institute of Technology, HIT Campus, Shenzhen University Town, Nanshan District, Shenzhen 518055, Guangdong, China

*Corresponding author. Info Building, HIT Campus, Shenzhen University Town, Nanshan District, Shenzhen 518055, China. E-mail: junjiechen@hit.edu.cn

Abstract

Antimicrobial peptides (AMPs) emerge as a type of promising therapeutic compounds that exhibit broad spectrum antimicrobial activity with high specificity and good tolerability. Natural AMPs usually need further rational design for improving antimicrobial activity and decreasing toxicity to human cells. Although several algorithms have been developed to optimize AMPs with desired properties, they explored the variations of AMPs in a discrete amino acid sequence space, usually suffering from low efficiency, lack diversity, and local optimum. In this work, we propose a novel directed evolution method, named PepZOO, for optimizing multi-properties of AMPs in a continuous representation space guided by multi-objective zeroth-order optimization. PepZOO projects AMPs from a discrete amino acid sequence space into continuous latent representation space by a variational autoencoder. Subsequently, the latent embeddings of prototype AMPs are taken as start points and iteratively updated according to the guidance of multi-objective zeroth-order optimization. Experimental results demonstrate PepZOO outperforms state-of-the-art methods on improving the multi-properties in terms of antimicrobial function, activity, toxicity, and binding affinity to the targets. Molecular docking and molecular dynamics simulations are further employed to validate the effectiveness of our method. Moreover, PepZOO can reveal important motifs which are required to maintain a particular property during the evolution by aligning the evolutionary sequences. PepZOO provides a novel research paradigm that optimizes AMPs by exploring property change instead of exploring sequence mutations, accelerating the discovery of potential therapeutic peptides.

Keywords: antimicrobial peptides; directed evolution; zeroth-order optimization

Introduction

Antimicrobial peptides (AMPs) are a type of small proteins that exhibit broad-spectrum antimicrobial activity with high specificity and good tolerability [1]. The emergence of drug resistance is becoming a major clinical challenge due to the single target antibiotics, long-term and extensive utilization [2, 3]. Unlike conventional antibiotics, AMPs can act on multiple targets on the plasma membrane and intracellular targets of pathogenic bacteria, demonstrating their potential as novel therapeutic candidates [4]. However, natural AMPs often have deficiencies, such as instability, short half-life, side effects, severe hemolytic activity, and proteolytic degradation [5]. Therefore, it is an urgent need to optimize the natural AMPs to overcome their shortcomings. Traditional chemical modification-based methods design new AMPs by adding or substituting amino acids, usually aiming at increasing the amphiphilicity or charge. However, such a process is costly and time-consuming due to the vast searching space. Thus, it is urgent to devise efficient and accurate *in silico* approaches to accelerate the discovery of novel AMP drugs.

In recent years, computational approaches for peptide design are rapidly emerging and have achieved significant success in the discovery of therapeutic peptides. The paradigms of existing methods are grouped as screening-based methods, *de novo* design methods, and evolutionary-based methods. The screening-based methods build classifiers to predict the properties of peptides, such as the toxicity or activity, by typically capturing the internal patterns of amino acids from large-scale databases. A series of methods based on quantitative structure-activity relationship models [6], traditional machine learning methods (iAMP-2L [7], ProFun-SOM [8], scCM [9], Lee et al. [10]), deep-learning methods (AMPlify [11], iAMPCN [12], iAMP-CA2L [13], TPpred-ATMV [14], iMFP-LG [15], MLBP [16], AMP Scanner Vr.2 [17], CFAGO [18], TPpred-SC [19], sAMPpred-GAT [20], KNIME [21], pLM4ACE [22], pLM4Alg [23], esm-AxP-GDL [24]), pre-trained protein language models [25, 26] were proposed to distinguish AMPs and non-AMPs. To discover new AMPs, these screening-based methods are usually employed to screen various large protein datasets. For instance, AMPlify [11] was used to mine the AMP-rich North American bullfrog (Rana [Lithobates] catesbeiana) genome for novel natural AMPs. Ma et al. [25] combined multiple natural language processing models to construct a unified pipeline for candidate AMP identification from human gut microbiome data. Nevertheless, these methods can only identify possible AMPs from existing databases and cannot generate novel AMPs.

Received: August 1, 2024. **Revised:** December 8, 2024. **Accepted:** December 27, 2024

The *de novo* design methods utilize generative models to learn the distribution of natural AMPs and generate novel potential active peptides that don't exist in nature. The widely used generative models, such as generative adversarial networks (GAN) (PepGAN [27], Enhancer-GAN [28], Forest-GAN [29], AMP-GAN [30], DeepImmuno [31], PAR-GAN [32], ProteinGAN [33], GAN-pep [34], Feedback GAN [35]) as well as variational autoencoders (VAE) (PepCVAE [36], PepVAE [37], HydrAMP [5], CLaSS [38]), have been employed for designing new AMPs. Recent progress in Transformer-based architectures has enabled the implementation of language models capable of generating text with human-like capabilities. Motivated by this success, protein language models trained on the protein space that generates *de novo* protein sequences following the principles of natural ones have been proposed, including ProtGPT2 [39], ProGen2 [40]. Nevertheless, these generative models have no precise control over the properties of the generated peptides. The candidate sequences are typically challenging to further optimize iteratively by this method.

In contrast, the evolutionary-based methods iteratively evolve a population of AMPs from the AMP prototypes and evaluate their fitness. The performance of these methods depend on the strategies of evolution. The genetic algorithms (GA) based approaches ([41–46]) generate new sequences by adding random mutations or recombination. Bayesian optimization (BO) [47] recommends the new AMP variants according to the distribution of existing AMPs so as to find AMPs with high properties efficiently. However, both GA and BO methods are used for searching in the discrete sequence space, resulting in low efficiency in the case of high discrete dimension. While the query-based molecule optimization [48] framework exploits latent representation learning and different sampling techniques to achieve an efficient search. Examples include the combined use of VAE and BO [49], VAE and Gaussian sampling [50], VAE and sampling guided by a predictor [51], VAE and evolutionary algorithms [52], VAE and random neighborhood sampling [48], deep reinforcement learning and a generative network [53, 54], and attribute-guided rejection sampling on a VAE [38].

Regarding the design of potential AMPs, optimizing the multiple properties of an AMP (e.g. overcoming the deficiencies of existing natural AMPs while enhancing their activity) can improve the success rate of generation. Compared with the exploration of whole peptide space by adding or substituting amino acids, the zeroth-order optimization algorithm is flexible to be integrated with multiple-objective optimization framework to directionally optimize the AMP in latent representation space, avoiding costly and time-consuming searching.

In this study, we proposed a multi-objective directed evolution method for AMP design, named PepZOO, to optimize multiple properties of the natural AMPs. PepZOO first projects the amino acid sequences of AMPs as latent embeddings by leveraging a pre-trained autoencoder model, and then searches the evolutionary direction by evaluating the multiple desired properties of the close AMP embeddings in the latent space. At last, the evolutionary direction is determined according to the zeroth-order optimization. PepZOO is flexible, as it can work with any predictor and does not require the predictor to be differentiable, just using feedback from a set of predictors to guide the optimization. Experimental results demonstrate PepZOO outperforms state-of-the-art methods on improving the properties of peptides in terms of antimicrobial function, activity, and toxicity. Furthermore, molecular docking as well as molecular dynamics (MD) simulations show the validity of PepZOO with a high success rate on the optimization of binding affinity. PepZOO is a novel paradigm in designing AMPs to accelerate the discovery of therapeutic AMPs and combat the issue of antimicrobial resistance.

Results

Overview of the proposed method PepZOO

PepZOO is a multi-objective optimization framework that directly searches peptide variants with desired properties in a peptide representation space guided by a set of property predictors and evaluation metrics. It consists of three modules (Fig. 1), including peptide representation module, property evaluation module, and directed searching module.

Peptide representation module projects peptides as embeddings in a latent representation space by leveraging an encoder-decoder model. The encoder Enc: $\mathbb{P}^L \to \mathbb{R}^d$ encodes a peptide $p = a_1, a_2, \cdots, a_L \in \mathbb{P}^L$ to a low-dimensional continuous real-valued representation of dimension d, denoted by an embedding vector $z = \mathrm{Enc}(p)$, where $a_i (i = 1, 2, \cdots, L)$ is one of 20 amino acids and $\mathbb{P}^L$ is a set of peptides with less than L amino acids. The decoder Dec: $\mathbb{R}^d \to \mathbb{P}^L$ decodes the latent representation z back to the peptide sequence p', denoted by $p' = \mathrm{Dec}(z)$. Property evaluation module can be a set of predictive models or scoring functions used to estimate desired properties of peptide variants generated by the decoder. For the generated peptide sequence p', PepZOO employs a set of separate predictive models $f_i(p')_{i=1}^m$ to evaluate the multiple properties to be optimized, where $f(\cdot)$ is denoted as a predictive model and m is the total number of predictive models. In this study, we are interested in the properties, such as antimicrobial function, activity, toxicity, and binding affinity. Directed searching module is to determine an evolutionary direction according to the property evaluation of n random samplings $z^{(1)}, z^{(2)}, \cdots, z^{(n)}$ near the embedding z in latent representation space by the property evaluation module $f(\mathrm{Dec}(z^{(i)}))$. Since the property evaluation module may not be differentiable, zeroth-order optimization algorithm is employed to calculate gradients of optimization direction, denoted by $\nabla grad(z) = \sum_{j=1}^m \sum_{i=1}^n \mathrm{ZOO}_j(f_j(\mathrm{Dec}(z^{(i)})))$. At last, the peptide embeddings are updated according to the zeroth-order gradients $z' = z + k\nabla grad(z)$ (k is the learning rate), and are fed into the decoder to generate novel peptides $p'' = \mathrm{Dec}(z')$. This optimization can be iteratively conducted, until the target properties of peptides satisfy the requirements.

Optimization of function and activity of antimicrobial peptides

We first illustrated the effectiveness of PepZOO to optimize the antimicrobial function and activity of peptides, comparing with three widely used methods: conditional VAE (CVAE) [55], PepCVAE [36], and HydrAMP [5]. Specifically, we evaluated the effectiveness of separately enhancing the antimicrobial function P_{AMP} and antimicrobial activity P_{MIC}, as well as simultaneously optimizing these two properties. To evaluate the improvement of compared methods, all peptides were divided into three cases according to their antimicrobial function P_{AMP} and antimicrobial activity P_{MIC}, where case 1 consists of peptides with high antimicrobial function (P_{AMP} >0.8) and high antimicrobial activity (P_{MIC} >0.5), case 2 consists of peptides with high antimicrobial function (P_{AMP} >0.8) and low antimicrobial activity ($P_{MIC} \leq 0.5$), and case 3 consists of the remaining peptides with low antimicrobial function ($P_{AMP} \leq 0.8$). For DRAMP dataset, case 1 has 5523 peptides, 8,453 peptides for case 2, and 3,147 peptides for case 3. For APD3 dataset, case1 has 687 peptides, 688 peptides for case 2, and 13 peptides for case 3, respectively. Here, PepZOO_AMP represents the model to

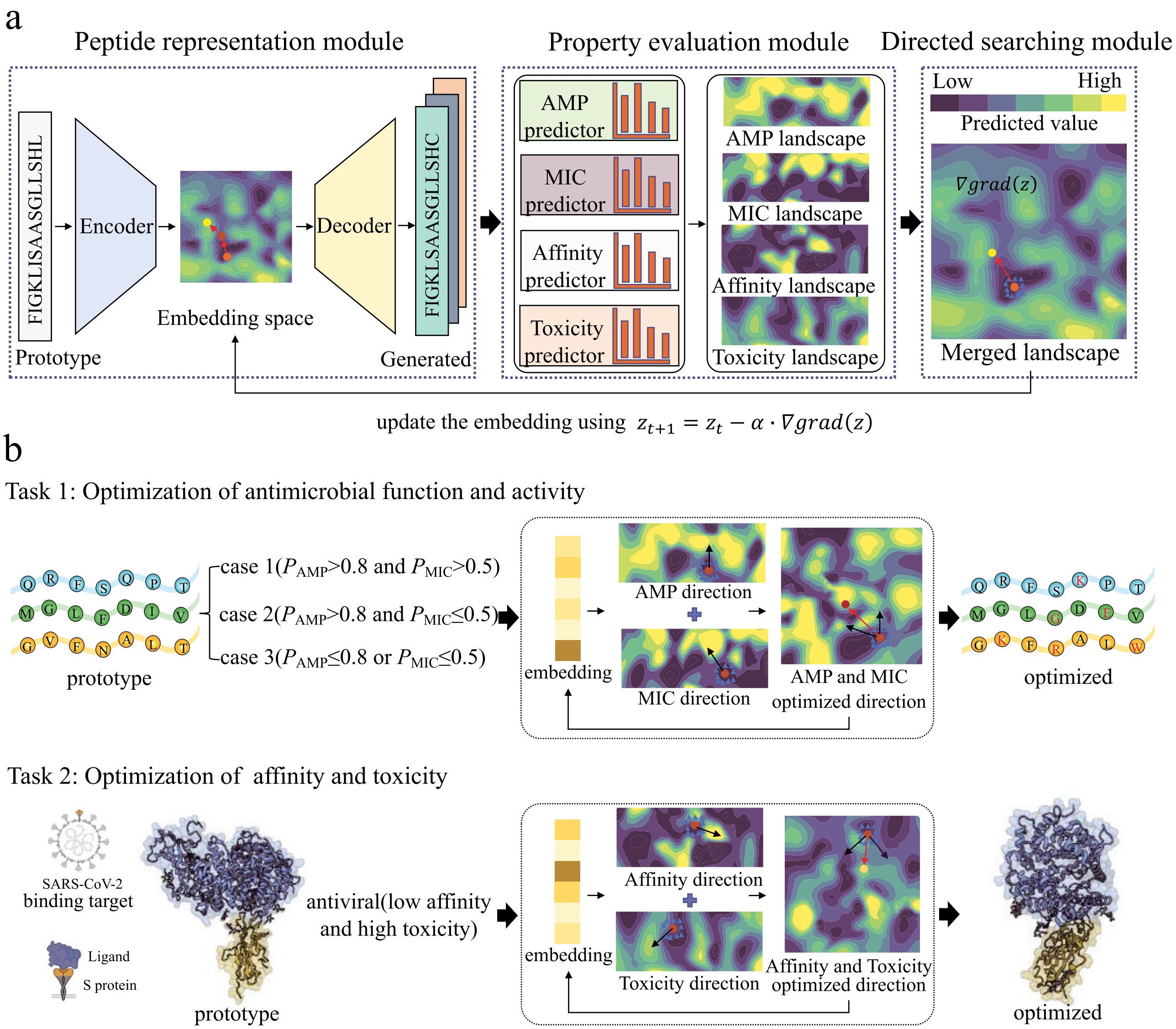

a
Peptide representation module
Property evaluation module
Directed searching module
FIGKLISAASGLLSHL
Prototype
Encoder
Embedding space
Decoder
FIGKLSAASGLLSHC
Generated
AMP predictor
MIC predictor
Affinity predictor
Toxicity predictor
AMP landscape
MIC landscape
Affinity landscape
Toxicity landscape
Low
High
Predicted value
∇grad(z)
Merged landscape
update the embedding using $z_{t+1} = z_t - \alpha \cdot \nabla grad(z)$
b
Task 1: Optimization of antimicrobial function and activity
Q R F S Q P T
M G L F D I V
G V F N A L T
prototype
case 1(P_{AMP}>0.8 and P_{MIC}>0.5)
case 2(P_{AMP}>0.8 and P_{MIC}≤0.5)
case 3(P_{AMP}≤0.8 or P_{MIC}≤0.5)
embedding
AMP direction
MIC direction
AMP and MIC optimized direction
Q R F S K P T
M G L G D F V
G K F R A L W
optimized
Task 2: Optimization of affinity and toxicity
SARS-CoV-2 binding target
Ligand
S protein
prototype
antiviral(low affinity and high toxicity)
embedding
Affinity direction
Toxicity direction
Affinity and Toxicity optimized direction
optimized

Figure 1. Illustration of the proposed PepZOO framework. (a) The architecture of PepZOO, which consists of three modules, including peptide representation module, property evaluation module, and directed searching module. (b) Two applications of PepZOO. The task 1 is to optimize the antimicrobial functionality of AMP prototypes, which are divided into three cases according to their antimicrobial function P_{AMP} and activity P_{MIC}. The task 2 is to optimize the structural property of prototype peptides against SARS-CoV-2 virus.

optimize antimicrobial function. PepZOO_MIC denotes the model to optimize the antimicrobial activity, where PepZOO_AMP_MIC refers to simultaneously optimizing both properties.

For the antimicrobial function optimization, PepZOO_AMP outperforms the other three methods (Fig. 2a). All of the methods can generate peptides with higher P_{AMP} than that of prototypes for all of three cases in both DRAMP and APD3 datasets, but the improvement ratio of PepZOO_AMP is the highest. Taking the results of the DRAMP dataset as an example (Fig. 2a). For case 1, although most of the P_{AMP} of prototypes are greater than 0.97, PepZOO can generate peptides with P_{AMP} greater than 0.99. For case 2, PepZOO generates peptides with P_{AMP} greater than 0.99 while other methods generate peptides with P_{AMP} greater than 0.97. For case 3, given prototypes with low P_{AMP}, PepZOO can generate peptides with P_{AMP} greater than 0.95, while other models have a relatively small increase in P_{AMP}. Furthermore, although P_{MIC} is not an optimization goal in this experimental setting, the P_{MIC} of generated peptides is also improved. The last two rows in Fig. 2a show the performance comparison between our method and compared methods on the APD3 dataset. The results on the APD3 are consistent with those on the DRAMP. In summary, PepZOO has superior performance in improving P_{AMP} compared with other models.

For the antimicrobial activity optimization, PepZOO_MIC also surpasses other methods (Fig. 2a). AMPs with high activity can exert antimicrobial effects at lower doses, which can fundamentally reduce the side effects caused by AMPs. Here, an AMP with a higher P_{MIC} indicates a higher probability to have $MIC \leq 10^{1.5} \simeq 32\mu g/mL$. Thus, a higher P_{MIC} is better for antimicrobial activity optimization. The first two rows in Fig. 2a present the experimental results of our method alongside comparisons with other methods on the DRAMP dataset. For case 1, all of the methods can generate peptides with better P_{MIC} than that of prototype, but PepZOO_MIC can achieve a higher improvement ratio compared to other methods. For case 2 and case 3, PepZOO_MIC can generate highly active AMPs with P_{MIC} greater than 0.8, while the other three methods often fail to optimize the antimicrobial activity of the prototype with most of peptides' P_{MIC} less than 0.2. Similarly,

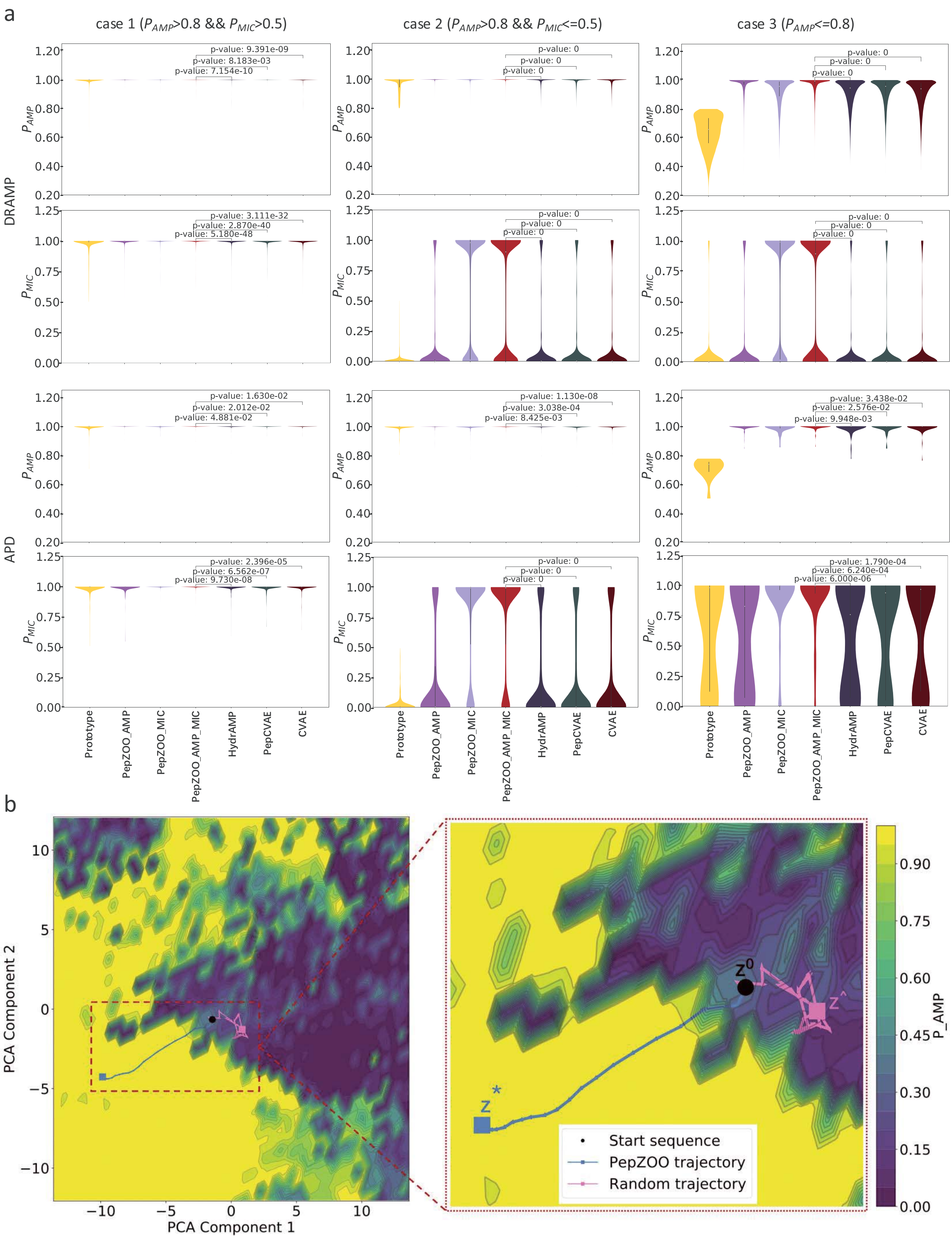


Figure 2. The optimization performance comparison between PepZOO and compared methods. (a) The performance comparison for all of three cases in DRAMP and APD3 datasets. (b) The optimization trajectory of PepZOO and random optimization. z^0 is the start sequence to optimize, z^* and $\hat{z}$ are the final optimized sequences by PepZOO and random method, respectively.

the results obtained from the APD3 dataset are consistent with those from the DRAMP dataset (Fig. 2a). Additionally, we found that P_{AMP} of the peptides generated by PepZOO_MIC has also been greatly improved in both DRAMP and APD3 datasets, which is consistent with the fact that there is a positive correlation between being AMP and high activity.

In order to make peptides satisfy multiple desired properties at the same time, we performed multi-objective optimization to simultaneously improve antimicrobial function and activity. Results of PepZOO_AMP_MIC in both DRAMP and APD3 show that both P_{AMP} and P_{MIC} of generated peptides have been greatly improved across all three cases, outperforming the separate optimization of antimicrobial function and activity. The first two rows in Fig. 2a show the experimental results of our method and compared methods on the DRAMP dataset. For prototypes in case 1 with P_{AMP} greater than 0.8 and P_{MIC} greater than 0.5, PepZOO_AMP_MIC can generate peptides with both P_{AMP} and P_{MIC} greater than 0.99, while PepZOO_AMP generates peptides with slightly lower P_{MIC}. For prototypes in case 2 with P_{AMP} greater than 0.8 and P_{MIC} less than 0.5, PepZOO_AMP_MIC can generate peptides with P_{AMP} greater than 0.99 and P_{MIC} greater than 0.8, while peptides generated by PepZOO_AMP get a much lower P_{MIC} with more than half of them less than 0.2. Surprisingly, PepZOO_MIC achieves the same performance as PepZOO_AMP_MIC for case 1 and 2. For prototypes in case 3 with P_{AMP} less than 0.8, PepZOO_AMP_MIC can generate peptides with P_{AMP} greater than 0.9 and P_{MIC} greater than 0.8, while PepZOO_AMP generates peptides with high P_{AMP} but low P_{MIC} and peptides generated by PepZOO_MIC get a lower P_{AMP} than peptides generated by PepZOO_AMP_MIC. Furthermore, the *t*-test method was employed to conduct a significance test on the experimental results. The *P*-values of between PepZOO_AMP_MIC and compared methods are all less than 0.05, indicating PepZOO_AMP_MIC significantly outperforms compared methods. The last two rows in Fig. 2a show the experimental results of our method in comparison with compared methods on the APD3 dataset. The results on the APD3 are also consistent with those on the DRAMP. These results demonstrate that PepZOO with multi-objective optimization not only achieves comparable improvement as separate optimization on one property, but also improves all desired properties simultaneously.

To further illustrate how PepZOO optimizes peptides in the representation space, we intuitively analyzed the PepZOO optimization trajectory in a 2D plane (Fig. 2b), comparing with the trajectory of a random optimization. We performed 20 iterations on the optimization processes of PepZOO and the random method. The optimization trajectory of PepZOO is guided toward a high P_{AMP} area, in contrast, the trajectory of a random process is an uncertain random walk. In the first five iterations, PepZOO was able to find a sequence with P_{AMP} greater than 0.9 in the search space, while P_{AMP} of the sequences generated by the random optimization process were all below 0.5, demonstrating that PepZOO has excellent directional optimization capabilities. In summary, these results showed that PepZOO can stably and efficiently explore peptides with better desired properties in the representation space.

Improvement of physicochemical properties and sequence novelty

The subtle balance between physicochemical properties of peptides and compositions of the amino acids determine the mode of action of AMP. Generally, AMPs have significantly larger isoelectric point, charge, hydrophobic ratio, and hydrophobic moment compared to those of non-AMPs. We estimated the distribution of the physicochemical properties of generated peptides, including charge, hydrophobic moment, hydrophobic ratio, and isoelectric point, as shown in Fig. 3a. The prototypes in case 1 are composed of AMPs with high P_{AMP} and high P_{MIC}, so they have relatively higher physicochemical properties than the whole AMP dataset. PepZOO preserves similar physicochemical properties except the hydrophobic moment, HydrAMP is also able to preserve these physicochemical properties except hydrophobic ratio, while CVAE and PepCVAE generate peptides with lower hydrophobic ratio and lower hydrophobic moment. For case 2, although its distribution of physicochemical properties is much lower than that of case 1, PepZOO improves the distribution of generated peptides to comparable with case 1. Nonetheless, other methods achieve limited improvements for all four physicochemical properties. As for case 3, the prototypes have the lowest distribution of physicochemical properties. PepZOO also notably improves the distribution of all four physicochemical properties, while other three competing methods fail to optimize this case.

Furthermore, we analyzed the novelty and improvement ratio during the optimization by PepZOO. Novelty is used to evaluate the differences between the optimized peptides and the prototypes. It is defined by the sum of Levenshtein distance between the prototype and corresponding optimized peptides divided by the number of optimized peptides. The larger novelty values indicate generated peptides have more mutations comparing the corresponding prototype sequences. The improvement ratio is the probability of successfully optimized peptides whose P_{AMP} and P_{MIC} are higher than the prototypes. As shown in Fig. 3b. For all three cases, the improvement ratio (red line) quickly reaches a stable value in several iterations, such as 20% of peptides in case 1 are improved by PepZOO in terms of MIC, about 30% for peptides in case 2, and 50% for peptides in case 3. Since peptides in case 1 whose P_{AMP} and P_{MIC} are already high, there is limited potential space for improvement. Whereas, PepZOO exhibits a higher relative growth in the improvement ratio in case 2 and case3, demonstrating it can achieve more significant improvements in the peptides with low P_{MIC}. However, the novelty (blue line) gradually increases with the number of iterations for all three cases, indicating that PepZOO attempts to explore different regions of latent space to design diverse peptides. For peptides in case 1 whose P_{AMP} and P_{MIC} are already high, PepZOO only needs to search for peptides that meet the desired properties around the latent space of prototypes. Compared with case 1, case 2 requires PepZOO to explore latent space farther away from prototypes to meet the requirement of high P_{MIC}. In order to simultaneously satisfy high P_{AMP} and P_{MIC}, given prototypes with low P_{AMP} and P_{MIC}, PepZOO must jump out of the current poor region to find a new region with high P_{AMP} and P_{MIC}. In summary, if only considering the improvement of MIC, PepZOO can optimize the prototypes in several iterations, while keeping high sequence identity. If considering to design novel peptides, more iterations are needed, and keeping high improvement ratio at the same time.

Optimization of antiviral peptides toward strong binding affinity to SARS-CoV-2 S protein and low toxicity to human organisms

We further illustrated the effectiveness of PepZOO to optimize protein structure by enhancing the binding affinity of existing antiviral peptides to SARS-CoV-2 S protein and decreasing their toxicity to human organisms. Binding affinity measures the ability of a peptide to bind to a target protein. The higher the binding

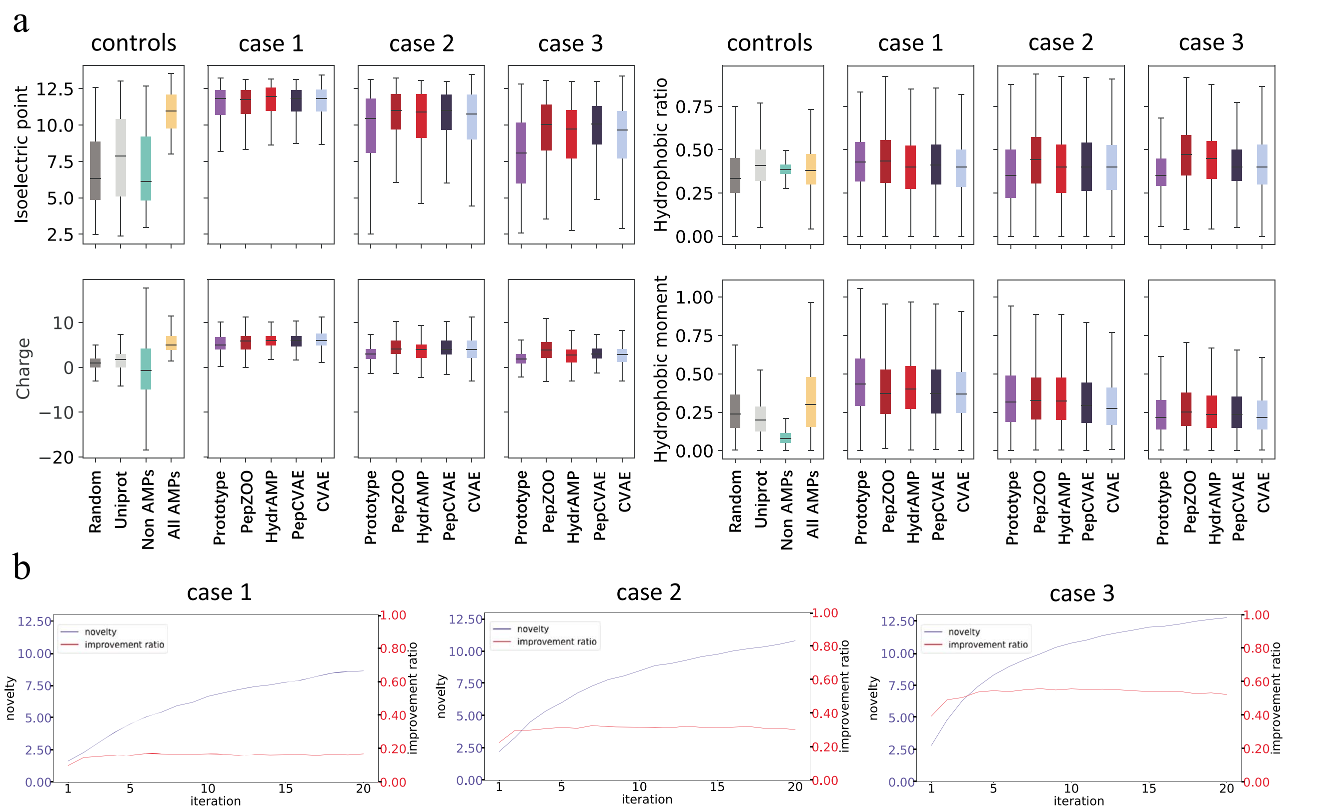
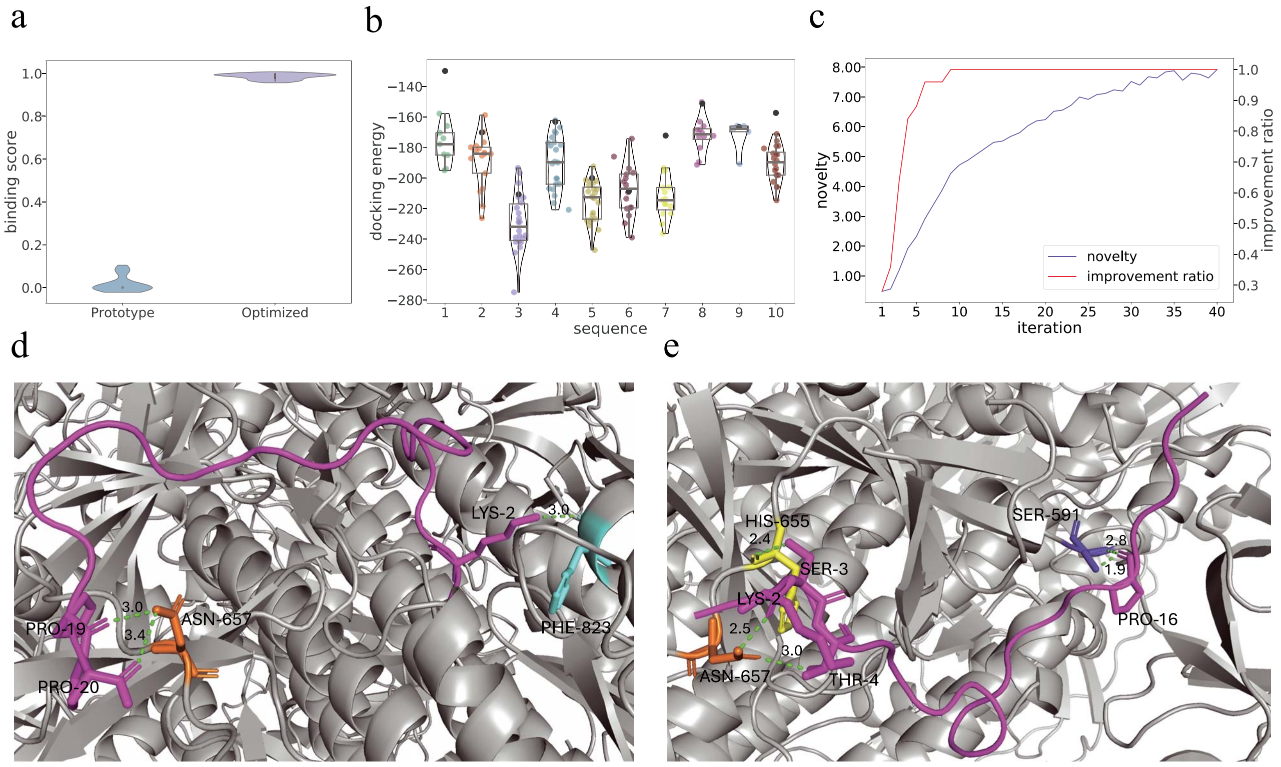


Figure 3. Comparison of improvement of physicochemical properties and sequence novelty. (a) Distributions of physicochemical properties (isoelectric point, charge, hydrophobic ratio, hydrophobic moment). (b) Generation performance in terms of novelty and improvement ratio.

affinity, the more stably the peptides bind to the target protein. Low toxicity represents the slight side effects. Toxicity is often the reason for the final failure of peptides in the clinical stage, peptides with high toxicity can cause serious side effects on the human body and therefore cannot be used to treat various human diseases.

We randomly selected 10 antiviral peptides from the antiviral dataset as prototypes due to the large time cost of MD simulations, and then performed 40 rounds of iterative optimization. The binding affinity of all peptides to SARS-CoV-2 S protein is evaluated by CAMP [56]. Most of the initial binding scores of prototypes are below 0.1, indicating that the prototypes have hardly binding ability to SARS-CoV-2 S protein. After optimization, their binding scores reach above 0.9, which means that optimized peptides could bind to SARS-CoV-2 S protein more stably (Fig. 4a). To further verify the effectiveness of optimized peptides, molecular docking was performed on those peptides with binding score greater than 0.9. The PDB file of SARS-CoV-2 S protein was downloaded from the PDB dataset while the PDB files of optimized peptides were generated by ESMFold [57]. Most peptides optimized by PepZOO need lower docking energy than the prototypes (Fig. 4b), indicating that they have better binding affinity to SARS-CoV-2 S protein. The trend of novelty and improvement ratio with the number of iterations is similar to the MIC optimization of AMPs. The novelty gradually increases with the number of iterations until the 35th iteration, while the improvement ratio reaches the highest value at the 10th iteration and remains until the end (Fig. 4c). In addition, we selected top 3 optimized peptides for all 10 prototypes according to the docking energy to run MD simulations. For 9 out of the 10 prototypes, we obtained optimized peptides with lower binding free energies, which means that these 9 optimized peptides have better ability to bind to SARS-CoV-2 S protein, the results of successfully optimized peptides and corresponding prototypes were shown in Table 1. For a case study, the binding sites of prototypes and optimized peptides to SARS-CoV-2 S protein are visualized in Fig. 4d and **e**, where the binding between prototype and the SARS-CoV-2 S protein have one hydrogen bond and one dihydrogen bond with docking energy of -170.092 kcal/mol. In contrast, the optimized peptide formed one hydrogen bond and two dihydrogen bonds with the SARS-CoV-2 S protein with docking energy of -218.429 kcal/mol, indicating that the optimized peptide could bind more tightly to the SARS-CoV-2 S protein.

We also attempted to simultaneously optimize the toxicity and binding affinity of existing antiviral peptides to SARS-CoV-2 S protein. The toxicity of peptides was predicted by Toxinpred3 [58], and a peptide with toxic score greater than or equal to 0.38 was considered toxic. We compared the improvements of binding score and toxic score in three different optimization strategies: optimizing binding affinity alone, optimizing toxicity alone, and optimizing both binding affinity and toxicity simultaneously. Taking the prototype sequence GVSGHGQHGVHG as an example, the improvement comparison of three different optimization strategies is shown in Fig. 5. For the strategy of optimizing binding affinity alone, the binding score of the optimized peptide quickly increases from nearby 0 to above 0.9, and then fluctuates around 0.9, but the toxic score remains above 0.4. Nonetheless, when optimizing toxicity alone, the toxic score of the optimized peptide dropped rapidly from over 0.9 to close to 0, while the binding score first increased and then decreased and stabilized at around 0.7 at last. When simultaneously optimizing binding affinity and toxicity, the binding scores reached the highest value in the 4th



Prototype (VKSTGRADDDLAVKTKYLPP)

Optimized (VKSTGYADDALAVPVPYTFP)

Figure 4. The results of binding affinity optimization. Distribution comparison of (a) binding scores and (b) docking energy between 10 prototype peptides and corresponding optimized peptides. The black dot represents the prototypes while other color dots in the same column represent corresponding optimized peptides. (c) The novelty and improvement ratio during optimization processes. The binding sites of SARS-CoV-2 S protein with (d) prototype and (e) optimized peptides. The green dotted line denotes the hydrogen bond, and the number above the green dotted line is the atomic distance between the two atoms forming a hydrogen bond.

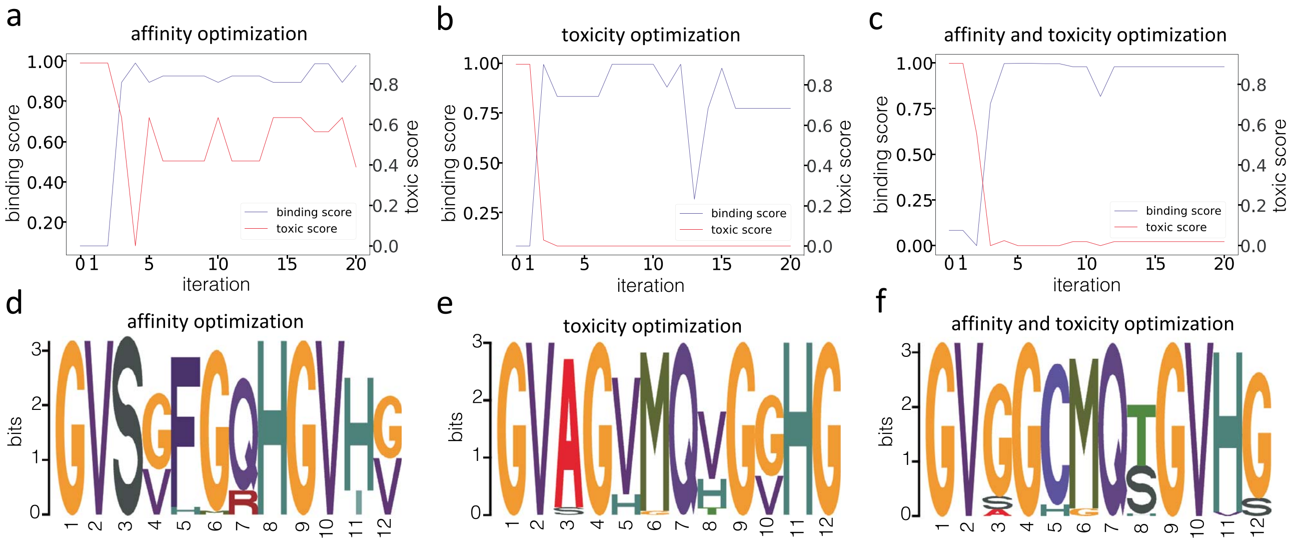


Figure 5. The optimization of binding affinity and toxicity. (a–c) The binding score and toxicity score during optimization by PepZOO when using binding affinity, toxicity, and binding affinity and toxicity as the optimization target, respectively. (d–f) The sequence logos of optimized peptides when using binding affinity, toxicity, and binding affinity and toxicity as the optimization target, respectively.

iteration, and finally stabilized above 0.9, while the toxic scores reached the lowest value in the 3rd iteration and then stabilized close to 0. These results demonstrate that when optimizing one property alone, the other property also can be optimized, but the improvement is limited. When simultaneously optimizing binding affinity and toxicity, the two target properties are both optimized significantly. To further illustrate the difference among these three optimization strategies, we aligned the optimized sequences to find the specific motifs (Fig. 5d–f). The 1st, 2nd, and 9th sites are kept conserved during optimization for all three optimization strategies. Besides, the peptides with low toxicity both in optimizing toxicity alone and simultaneously optimizing binding affinity and toxicity have more conserved sites, including 4th, 6th, 7th, 11th sites. Compared with the peptides with high toxicity in

Table 1. The binding free energy of MD simulations results of prototypes and corresponding optimized peptides.

Type	Sequence	Binding free energy (kcal/mol) ↓
Prototype	EQCREEEDDR	+43.47
Optimized	EQFRLELSAR	−17.59
Prototype	VKSTGRADDDLAVKTKYLPP	+0.47
Optimized	VKSTGYADDALAVPVPYTFP	−15.74
Prototype	GWMSKIASGIGTFLSGV---QQG	+19.81
Optimized	GKMTWFAEGCGVPISGHYTHQNM	−49.92
Prototype	LLKELWTKIKGAGKAVLGKIKGLL	+3.30
Optimized	ELNDTKGISHGGAKAGGALLHAHV	−52.13
Prototype	GVSGHGQHG--VHG	−1.85
Optimized	GFSHFGRFGVPVHT	−21.08
Prototype	GLLSGILNTAGGLLGNLIGSLSN	+3.19
Optimized	GLKTVFIPHGGVLHSITTSGEGH	−38.00
Prototype	GIADILKGLL	+15.69
Optimized	GFSEFNKGLL	−5.30
Prototype	LLGGLLQSLL---	+0.50
Optimized	LLGGLLIGLETIN	−11.06
Prototype	GVVDTLKNLLMGLL-	−3.82
Optimized	GVVAILKPHHGGLLF	−12.25
Prototype	NRILPTLIGPL	−0.07
Optimized	LR-PFYVIHM	+7.05

Note: ↓ means a smaller value is better on this metric. The amino acids represented by orange letters have changed after optimization, the amino acids represented by blue letters are the amino acids added after optimization, and the amino acids represented by red letters are the amino acids that were deleted after optimization.

optimizing binding affinity alone, the 3rd, 5th, 6th sites are totally different, which means these sites may have a critical impact on toxicity. Overall, all the above results demonstrate that PepZOO can effectively optimize the binding affinity of existing antiviral peptides to the target protein while reducing their toxicity and provide promising drug candidates for the drug discovery process.

Conclusion

In this study, we proposed a novel directed evolution method, named PepZOO, for optimizing multi-properties of peptides. PepZOO projects AMP sequences into a continuous latent representation space by a VAE and searches the evolutionary direction guided by multi-objective zeroth-order optimization. Experimental results on two tasks, including the optimization of antimicrobial function and activity and the optimization of affinity and toxicity, showed that PepZOO outperforms the state-of-the-art methods, especially for those cases with poor properties. In summary, PepZOO can effectively optimize various objectives and constraints based on initial peptide sequences, indicating that it can be applied to the actual peptide design process and propose new peptides with good target binding ability. PepZOO can serve as a practical tool for peptide optimization and accelerate discovery of peptide drugs.

Although our study demonstrates outstanding results, there are two primary limitations to be considered in future work. First, the peptide sequence length is restricted to 25 amino acids. This threshold facilitates the success rate of subsequent laboratory synthesis, but it concurrently constrains the sequence space that the peptide representation model can learn. Moving forward, we plan to enhance our model to accommodate longer peptide sequences, which will enable us to better assess their potential binding affinity and therapeutic applications. Given the remarkable performance of large language models in natural language processing, future work could leverage protein language models trained on extensive protein sequence data to enhance the peptide representation and generation capabilities. Second, the zeroth-order optimization introduces randomness, which may limit the model's performance. The zeroth-order gradient is derived from random perturbations around the peptide embeddings, potentially affecting the model's convergence speed.

Materials and methods

Datasets

We evaluated our proposed method on two widely used datasets, Data Repository of Antimicrobial Peptides (DRAMP) [59], which currently encompasses a collection of 22 528 entries, and APD3 with a total number of 3167 unique sequences. Since the AMPs with less than 25 amino acids are more feasible to peptide synthesis and AMPs with a length of less than or equal to 25 amino acids account for more than 70% of the entire dataset, peptide sequences longer than 25 amino acids were filtered out, as a result that 17 123 AMPs were selected from DRAMP and a total number of 1388 entries were selected from APD3. We used these benchmark datasets to test the performance of PepZOO as well as compared methods. AMP is a general term for peptides with different functions. Generally speaking, AMPs can be divided into different categories according to their functions, such as antibacterial peptides, antiviral peptides, anticancer peptides, etc. For the antimicrobial function and activity optimization task, we use all AMPs as prototypes to evaluate our proposed method and compared methods. Although Müller et al. [60] proposed a method to design peptides against the SARS-CoV-2 S protein and obtain some inhibitors toward the SARS-CoV-2 S protein, the length of these inhibitors exceed 25 amino acids, which limits our ability to incorporate them into our analysis at this time. Besides, due to the huge amount of running time for MD simulation, we randomly selected 10 antiviral peptides with high toxicity and low binding affinity to SARS-CoV-2 S protein as the prototypes for the binding affinity and toxicity optimization task.

Peptide representation module

Peptide representation module projects peptides as latent embeddings. Since a peptide is a discrete string of amino acids, the gradient optimization can't efficiently search for peptides with desired properties in discrete amino acid combination space. Therefore, peptides have to be mapped into a continuous representation space. The peptide representation module can be any encoder-decoder models, which satisfy two conditions: (i) peptides with similar properties are projected as embeddings closely in the representation space by the encoder; (ii) these close embeddings in latent space can be reconstructed into peptides with similar properties by the decoder.

In this study, we employed a CVAE model, named HydrAMP [5] as the peptide representation model. HydrAMP incorporates two optimization objectives: Jacobian disentanglement regularization and reconstruction regularization. The Jacobian disentanglement regularization encourages an orthogonal decoupling between the latent variable z and two properties, antimicrobial function and activity. The reconstruction regularization focuses on training the VAE model to generate valid peptide structures by accurately reconstructing input peptides. HydrAMP learns the spatial distribution of peptide sequences and decouples the representation of peptides from their properties. It can not only satisfy the aforementioned two conditions but also ensure that the generated peptides adhere to specific criteria.

Property evaluation module

Property evaluation module estimates interested properties of generated peptides. In this study, we focus on the AMPs with high activity, high affinity, and low toxicity. Thus, we employed an AMP predictor, an MIC predictor, a toxicity predictor, and a binding affinity predictor to estimate the properties of each generated novel peptide.

The ESM-2 [61] can learn evolutionary information [62] in large-scale protein sequences, enabling it to achieve excellent performance in downstream tasks. Therefore, the antimicrobial function of peptides is estimated through the probability of being AMPs (P_{AMP}) by an AMP predictor developed by Cordoves et al. [24], which is a graph deep learning (GDL) architecture based on ESM-2 and ESMFold [61] trained on a comprehensive dataset. Its experimental results show that the performance of this model is optimal with an accuracy of 0.97, outperforming the other competing models. The activity of peptides (P_{MIC}) is predicted by an MIC predictor proposed by Szymczak et al. [5], which estimates whether AMPs are highly active with $MIC \leq 10^{1.5} \simeq 32\mu g/mL$. This model consists of CNN and LSTM architecture and achieves an accuracy of 0.942 on experimentally validated MIC data, outperforming the other competing models. The toxicity of peptides is predicted by Toxinpred3 [58], which is an integrated method by building multiple decision trees and combining their prediction results, achieving an AUROC of 0.98 and an MCC of 0.81. In order to predict the binding affinity of a peptide with its target protein, CAMP [56] is employed, which is a state-of-the-art method on binary peptide-protein interaction prediction via CNN and self-attention layers to extract features of peptides and proteins, respectively.

Directed searching module

Directed searching module determines an evolutionary direction according to the feedback of property evaluation module. Since the property evaluation module may not be differentiable, a zeroth-order optimization algorithm is employed to calculate gradients of optimization direction. Because the properties are predicted based on discrete amino acids sequences, which can not directly feedback the optimization gradient about the embedding of peptides. Therefore, zeroth-order optimization is utilized to achieve gradients about the embedding of peptides according to the feedback of the property evaluation module.

Zeroth-order optimization does not require the differentiable evaluation function, but it defines a substitute based on sampling and difference, which is called zeroth-order gradient. It can be formulated as follows:

$$\widetilde{\nabla} f(p) = \mathbb{E}_{\mu \sim \mathcal{N}(\mu)} \left[\frac{f(Dec(\mathrm{z} + \varepsilon \cdot \mu)) - f(Dec(\mathrm{z}))}{\varepsilon} \cdot \mu \right] \tag{1}$$

where ε is a small positive number, $\mathcal{N}(\mu)$ is a normal distribution with a mean value of 0 and a covariance matrix as the unit matrix.

Multi-objective optimization is a crucial component in the directed searching module, particularly when dealing with the optimization of multiple peptide properties. In this study, different properties of peptides such as antimicrobial function, antimicrobial activity, binding affinity, and toxicity are considered simultaneously. The goal is to find an optimal balance among these properties to achieve the best overall performance of peptides. The multi-objective optimization approach integrated into the directed searching module enables the simultaneous optimization of multiple peptide properties, providing a balanced and effective solution to peptide design. Given a peptide p, firstly its latent embedding vector $\mathrm{z} = \mathrm{Enc}(p)$ is projected by an Encoder. Next, Q perturbation vectors $\{\mu_q^{(t)}\}_{q=1}^{Q}$ from normal distribution are randomly selected. Then the zeroth-order gradient at the t-th iteration is calculated by:

$$\widetilde{\nabla} grad^{(\mathrm{t})}(\mathrm{z}) = \frac{d}{\beta \cdot \mathrm{Q}} \sum_{q=1}^{\mathrm{Q}} \sum_{\mathrm{i}=1}^{m} \omega_{\mathrm{i}} \cdot \left[f_{\mathrm{i}}(Dec(\mathrm{z}^{(\mathrm{t})} + \beta \cdot \mu_q^{(\mathrm{t})})) - f_{\mathrm{i}}(Dec(\mathrm{z}^{(\mathrm{t})})) \right] \cdot \mu_q^{(\mathrm{t})} \tag{2}$$

where d is the dimension of the latent space learned by the Encoder, $\beta > 0$ is a smoothing parameter used to perturb the embedding vector $\mathrm{z}^{(\mathrm{t})}$, t is the number of iterations, $Dec(\cdot)$ is the Decoder, f_{i} denotes property predictors, specifically in this study, f_1 denotes AMP predictor, f_2 denotes MIC predictor, f_3 denotes binding affinity predictor and f_4 denotes toxicity predictor, and ω_{i} is the weight of each predictor. By setting different weights for different predictors, our method is able to optimize multiple desired properties simultaneously and weigh the importance of different properties. Finally, we use the zeroth-order gradient to update the embedding vector $\mathrm{z}^{(\mathrm{t})}$ by:

$$\mathrm{z}^{(\mathrm{t}+1)} = \mathrm{z}^{(\mathrm{t})} - \alpha \cdot \widetilde{\nabla} grad(\mathrm{z}^{(\mathrm{t})}) \tag{3}$$

where α is the learning rate. In this study, d is set to 64, β is set to 0.5, Q is set to 32, m is set to 4, and α_{t} is set to 0.05.

Molecular docking

Molecular docking predicts the binding affinity of ligands to receptor proteins by simulating the interaction between a small molecule and a protein at the atomic level, subsequently enabling researchers to study the behavior of small molecules within the binding site of a target protein and understand the fundamental biochemical process underlying this interaction.

In this work, molecular docking of peptides to the SARS-CoV-2 S protein is predicted by HPEPDOCK [63], which is a web server for global peptide–protein docking based on a hierarchical algorithm. A comprehensive evaluation [64] among 14 docking programs on peptide-protein complexes demonstrates that HPEPDOCK has the best success rate and computational efficiency in global docking. For binding affinity optimization, these optimized peptides with binding score given by CAMP greater than or equal to 0.9 were selected to dock with the SARS-CoV-2 S protein. Then the top three conformations generated by HPEPDOCK were selected for MD simulations.

Molecular dynamics simulations

MD simulations of the peptide-protein complexes were performed using the GROMACS 2020.7 [65] with CUDA support. The top three conformations of the docking results generated by HPEPDOCK were selected as the initial peptide-protein complexes to perform MD simulations.

MD simulations were carried out with Amber ff99SB-ILDN force field [66], which is a force field improved by modifying the parameters of side chain based on Amber ff99SB [67]. Firstly, the structure files of the docked complexes were converted into the topology files in GROMACS format by the pdb2gmx module. Then a dodecahedron box was constructed by editconf module and the peptide-protein complex was placed in its center, ensuring a minimum distance of 1.2 nm between the complex and the edges of the box. The dodecahedron box was filled with TIP3P water molecules, and the counter ions (Na^+ and Cl^-) were added to the box to neutralize the total charge. After that, several pre-equilibrium steps for energy minimization using the steepest descent minimization method under the vacuum and solvated environment were performed, respectively. The Particle Mesh Ewald [68] approach was used to calculate the long-range electrostatic interactions under the periodic boundary conditions, with a cutoff of 1.5 nm. Van der Waals non-bonded interactions were calculated by cutoff scheme with a cutoff of 1.5 nm. Then, NVT and NPT pre-equilibrium were performed to make the system reach a proper temperature at 310K and a pressure at 1 atm. Finally, the formal MD simulations for all of the peptide-protein complexes were performed with a simulation time of 200 ns, and the integration step was set to 2 fs.

Key Points

- This study proposed a novel directed evolution method, named PepZOO, for optimizing multi-properties of AMPs in a continuous representation space guided by a multi-objective zeroth-order optimization.
- PepZOO can simultaneously optimize the multiple desired properties of natural AMPs, achieving comparable improvements with optimizing each property alone.
- PepZOO can reveal important motifs which are required to maintain a particular property during the evolution by aligning the evolutionary sequences.

Funding

This work was supported by the National Natural Science Foundation of China (62102118, 82273890), the Basic and Applied Basic Research Foundation of Guangdong Province (2021A1515220115), the Project of Educational Commission of Guangdong Province of China (2022KTSCX210), Shenzhen Science and Technology Program (JCYJ20230807094318038), Shenzhen Colleges and Universities Stable Support Program (GXWD20220811170504001, GXWD20231130103401001), Futian Healthcare Research Project (FTWS2023091).

Data availability

All the datasets used in this study and all source codes of the PepZOO algorithm have been deposited at https://github.com/chen-bioinfo/PepZOO.

References

1. Lei J, Sun L, Huang S. *et al*. The antimicrobial peptides and their potential clinical applications. *Am J Transl Res* 2019;**11**:3919–31.
2. Solomon SL, Oliver KB. Antibiotic resistance threats in the United States: stepping back from the brink. *Am Fam Physician* 2014;**89**:938–41.
3. Frieri M, Kumar K, Boutin A. Antibiotic resistance. *J Infect Public Health* 2017;**10**:369–78. https://doi.org/10.1016/j.jiph.2016.08.007.
4. Zhang Q, Yan Z, Meng Y. *et al*. Antimicrobial peptides: mechanism of action, activity and clinical potential. *Mil Med Res* 2021;**8**:48. https://doi.org/10.1186/s40779-021-00343-2.
5. Szymczak P, Możejko M, Grzegorzek T. *et al*. Discovering highly potent antimicrobial peptides with deep generative model HydrAMP. *Nat Commun* 2023;**14**:1453. https://doi.org/10.1038/s41467-023-36994-z.
6. Cardoso MH, Orozco R, Rezende SB. *et al*. Computer-aided design of antimicrobial peptides: are we generating effective drug candidates? *Front Microbiol* 2020;**10**:3097. https://doi.org/10.3389/fmicb.2019.03097.
7. Xuan Xiao P, Wang W-ZL, Jia J-H. *et al*. iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal Biochem* 2013;**436**:168–77. https://doi.org/10.1016/j.ab.2013.01.019.
8. Shao JY, Chen JJ, Liu B. ProFun-SOM: protein function prediction for specific ontology based on multiple sequence alignment reconstruction. *IEEE Trans Neural Netw Learn Syst*. https://doi.org/10.1109/TNNLS.2024.3419250.
9. Fang Y, Chen J, He W. *et al*. Integrating large-scale single-cell RNA sequencing in central nervous system disease using self-supervised contrastive learning. *Commun Biol* **7**:1107. https://doi.org/10.1038/s42003-024-06813-2.
10. Lee EY, Fulan BM, Wong GCL. *et al*. Mapping membrane activity in undiscovered peptide sequence space using machine learning. *Proc Natl Acad Sci* 2016;**113**:13588–93. https://doi.org/10.1073/pnas.1609893113.
11. Li C, Darcy Sutherland S, Hammond A. *et al*. AMPlify: attentive deep learning model for discovery of novel antimicrobial peptides effective against WHO priority pathogens. *BMC Genomics* 2022;**23**:1–15.
12. Jing X, Li F, Li C. *et al*. iAMPCN: a deep-learning approach for identifying antimicrobial peptides and their functional activities. *Brief Bioinform* 2023;**24**:bbad240.
13. Xiao X, Shao Y-T, Cheng X. *et al*. iAMP-CA2L: a new CNN-BiLSTM-SVM classifier based on cellular automata image for identifying antimicrobial peptides and their functional types. *Brief Bioinform* 2021;**22**:bbab209. https://doi.org/10.1093/bib/bbab209.
14. Yan K, Lv H, Guo Y. *et al*. TPpred-ATMV: therapeutic peptide prediction by adaptive multi-view tensor learning model. *Bioinformatics* 2022;**38**:2712–8. https://doi.org/10.1093/bioinformatics/btac200.
15. Jiawei L, Kejuan Z, Junjie C. *et al*. iMFP-LG: identification of novel multi-functional peptides by using protein language models and graph-based deep learning. *Genomics Proteomics Bioinformatics* 2024;**22**:qzae084. https://doi.org/10.1093/gpbjnl/qzae084.
16. Tang W, Dai R, Yan W. *et al*. Identifying multi-functional bioactive peptide functions using multi-label deep learning. *Brief Bioinform* 2022;**23**:bbab414. https://doi.org/10.1093/bib/bbab414.
17. Veltri D, Kamath U, Shehu A. Deep learning improves antimicrobial peptide recognition. *Bioinformatics* 2018;**34**:2740–7. https://doi.org/10.1093/bioinformatics/bty179.
18. Zhourun W, Guo M, Jin X. *et al*. CFAGO: cross-fusion of network and attributes based on attention mechanism for protein function prediction. *Bioinformatics* 2023;**39**:btad123.
19. Yan K, Lv H, Shao J. *et al*. TPpred-SC: multi-functional therapeutic peptide prediction based on multi-label supervised contrastive learning. *Sci China Inf Sci* 2024;**67**:212105. https://doi.org/10.1007/s11432-024-4147-8.
20. Yan K, Lv H, Guo Y. *et al*. sAMPpred-GAT: prediction of antimicrobial peptide by graph attention network and predicted peptide structure. *Bioinformatics* 2023;**39**:btac715. https://doi.org/10.1093/bioinformatics/btac715.
21. Martínez-Mauricio KL, García-Jacas CR, Cordoves-Delgado G. Examining evolutionary scale modeling-derived different-dimensional embeddings in the antimicrobial peptide classification through a KNIME workflow. *Protein Sci* 2024;**33**:e4928. https://doi.org/10.1002/pro.4928.

22. Zhenjiao D, Ding X, Hsu W. *et al*. pLM4ACE: a protein language model based predictor for antihypertensive peptide screening. *Food Chem* 2024;**431**:137162. https://doi.org/10.1016/j.foodchem.2023.137162.
23. Zhenjiao D, Yixiang X, Liu C. *et al*. PLM4Alg: protein language model-based predictors for allergenic proteins and peptides. *J Agric Food Chem* 2024;**72**:752–60. https://doi.org/10.1021/acs.jafc.3c07143.
24. Cordoves-Delgado G, García-Jacas CR. Predicting antimicrobial peptides using ESMFold-predicted structures and ESM-2-based amino acid features with graph deep learning. *J Chem Inf Model* 2024;**64**:4310–21. https://doi.org/10.1021/acs.jcim.3c02061.
25. 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. https://doi.org/10.1038/s41587-022-01226-0.
26. Zhang Y, Lin J, Zhao L. *et al*. A novel antibacterial peptide recognition algorithm based on BERT. *Brief Bioinform* 2021;**22**:bbab200. https://doi.org/10.1093/bib/bbab200.
27. 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. https://doi.org/10.1021/acsomega.0c02088.
28. Li J, Xiao L, Luo J. *et al*. High-activity enhancer generation based on feedback GAN with domain constraint and curriculum learning. In: *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, Istanbul, Turkiye, pp. 2065–70. Piscataway, NJ, USA: IEEE, 2023.
29. Chen J, Li J, Song C. *et al*. Discriminative forests improve generative diversity for generative adversarial networks. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. **38**, pp. 11338–45. Palo Alto, California, USA: AAAI, 2024.
30. Ferrell JB, Remington JM, Van Oort CM. *et al*. A generative approach toward precision antimicrobial peptide design. *BioRxiv* 2020;2020–10.
31. Li G, Balaji Iyer VB, Prasath S. *et al*. DeepImmuno: deep learning-empowered prediction and generation of immunogenic peptides for T-cell immunity. *Brief Bioinform* 2021;**22**:bbab160. https://doi.org/10.1093/bib/bbab160.
32. Chen J, Wang WH, Gao H. *et al*. PAR-GAN: improving the generalization of generative adversarial networks against membership inference attacks. In: *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 127–37. New York, USA: Association for Computing Machinery (ACM), 2021.
33. Repecka D, Jauniskis V, Karpus L. *et al*. Expanding functional protein sequence spaces using generative adversarial networks. *Nat Mach Intell* 2021;**3**:324–33. https://doi.org/10.1038/s42256-021-00310-5.
34. Lin T-T, Yang L-Y, Wang C-T. *et al*. Discovering novel antimicrobial peptides in generative adversarial network. *bioRxiv* 2021; 2021–11.
35. Gupta A, Zou J. Feedback GAN for DNA optimizes protein functions. *Nat Mach Intell* 2019;**1**:105–11. https://doi.org/10.1038/s42256-019-0017-4.
36. Das P, Wadhawan K, Chang O. *et al*. PepCVAE: semi-supervised targeted design of antimicrobial peptide sequences arXiv preprint arXiv:1810.07743. 2018.
37. Dean SN, Jerome Anthony E, Alvarez DZ. *et al*. PepVAE: variational autoencoder framework for antimicrobial peptide generation and activity prediction. *Front Microbiol* 2021;**12**:725727. https://doi.org/10.3389/fmicb.2021.725727.
38. 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. https://doi.org/10.1038/s41551-021-00689-x.
39. Ferruz N, Schmidt S, Höcker B. ProtGPT2 is a deep unsupervised language model for protein design. *Nat Commun* 2022;**13**:4348. https://doi.org/10.1038/s41467-022-32007-7.
40. Nijkamp E, Ruffolo JA, Weinstein EN. *et al*. ProGen2: exploring the boundaries of protein language models. *Cell Syst* 2022;**14**:968–978.e3. https://doi.org/10.1016/j.cels.2023.10.002.
41. Liu Y, Zhang X, Liu Y. *et al*. Evolutionary multi-objective optimization in searching for various antimicrobial peptides [feature]. *IEEE Comp Intell Mag* 2023;**18**:31–45. https://doi.org/10.1109/MCI.2023.3245731.
42. Yoshida M, Hinkley T, Tsuda S. *et al*. Using evolutionary algorithms and machine learning to explore sequence space for the discovery of antimicrobial peptides. *Chem* 2018;**4**:533–43. https://doi.org/10.1016/j.chempr.2018.01.005.
43. Porto WF, Irazazabal LSF, Alves ESF. *et al*. In silico optimization of a guava antimicrobial peptide enables combinatorial exploration for peptide design. *Nat Commun* 2018;**9**:1490.
44. Fjell CD, Jenssen H, Cheung WA. *et al*. Optimization of antibacterial peptides by genetic algorithms and cheminformatics. *Chem Biol Drug Des* 2011;**77**:48–56. https://doi.org/10.1111/j.1747-0285.2010.01044.x.
45. Yuan Y, Pei J, Lai L. LigBuilder 2: a practical de novo drug design approach. *J Chem Inf Model* 2011;**51**:1083–91. https://doi.org/10.1021/ci100350u.
46. Reutlinger M, Rodrigues T, Schneider P. *et al*. Multi-objective molecular de novo design by adaptive fragment prioritization. *Angewandte Chemie* 2014;**53**:4244–8. https://doi.org/10.1002/anie.201310864.
47. Korovina K, Xu S, Kandasamy K. *et al*. ChemBO: Bayesian optimization of small organic molecules with synthesizable recommendations. In: *International Conference on Artificial Intelligence and Statistics*. Proceedings of Machine Learning Research (PMLR), 2019. [Online].
48. Hoffman SC, Chenthamarakshan V, Wadhawan K. *et al*. Optimizing molecules using efficient queries from property evaluations. *Nat Mach Intell* 2020;**4**:21–31. https://doi.org/10.1038/s42256-021-00422-y.
49. Griffiths R-R, Lobato JMH. Constrained Bayesian optimization for automatic chemical design using variational autoencoders. *Chem Sci* 2019;**11**:577–86. https://doi.org/10.1039/C9SC04026A.
50. Boitreaud J, Mallet V, Oliver CG. *et al*. OptiMol: optimization of binding affinities in chemical space for drug discovery. *bioRxiv* 2020.
51. Tianfan F, Xiao C, Sun J. Core: automatic molecule optimization using copy & refine strategy *ArXiv*, abs/1912.05910. 2019.
52. Winter R, Montanari F, Steffen A. *et al*. Efficient multi-objective molecular optimization in a continuous latent space. *Chem Sci* 2019;**10**:8016–24. https://doi.org/10.1039/C9SC01928F.
53. Olivecrona M, Blaschke T, Engkvist O. *et al*. Molecular de-novo design through deep reinforcement learning. *J Cheminform* 2017;**9**:1–14. https://doi.org/10.1186/s13321-017-0235-x.
54. Sánchez-Lengeling B, Outeiral C, Guimaraes GL. *et al*. optimizing distributions over molecular space. An objective-reinforced generative adversarial network for inverse-design chemistry (ORGANIC). *ChemRxiv*2017.
55. Zhiting H, Yang Z, Liang X. *et al*. Toward controlled generation of text. In: *International Conference on Machine Learning*. pp. 1587–96, Sydney, Australia: Proceedings of Machine Learning Research (PMLR), 2017. [Online].

56. Lei Y, Li S, Liu Z. *et al.* A deep-learning framework for multi-level peptide–protein interaction prediction. *Nat Commun* 2021;**12**:5465. https://doi.org/10.1038/s41467-021-25772-4.
57. Lin Z, Akin H, Rao R. *et al.* Language models of protein sequences at the scale of evolution enable accurate structure prediction. *BioRxiv* 2022. 500902, 2022.
58. Rathore AS, Arora A, Choudhury SPS. *et al.* ToxinPred 3.0: an improved method for predicting the toxicity of peptides. *bioRxiv* 2023;2023–08.
59. Shi G, Kang X, Dong F. *et al.* DRAMP 3.0: an enhanced comprehensive data repository of antimicrobial peptides. *Nucleic Acids Res* 2022;**50**:D488–96. https://doi.org/10.1093/nar/gkab651.
60. Desiree Schütz A, Ruiz Blanco B, Jan Münch A. *et al.* Peptide and peptide-based inhibitors of SARS-CoV-2 entry. *Adv Drug Deliv Rev* 2020;**167**:47–65. https://doi.org/10.1016/j.addr.2020.11.007.
61. Lin Z, Akin H, Rao R. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 2023;**379**:1123–30. https://doi.org/10.1126/science.ade2574.
62. Li J, Zhourun W, Lin W. *et al.* iEnhancer-ELM: improve enhancer identification by extracting position-related multiscale contextual information based on enhancer language models . *Bioinform Adv* 2023;**3**:vbad043.
63. Zhou P, Jin B, Li H. *et al.* HPEPDOCK: a web server for blind peptide–protein docking based on a hierarchical algorithm. *Nucleic Acids Res* 2018;**46**:W443–50. https://doi.org/10.1093/nar/gky357.
64. Weng G, Gao J, Wang Z. *et al.* Comprehensive evaluation of fourteen docking programs on protein–peptide complexes. *J Chem Theory Comput* 2020;**16**:3959–69. https://doi.org/10.1021/acs.jctc.9b01208.
65. Van Der Spoel D, Lindahl E, Hess B. *et al.* GROMACS: fast, flexible, and free. *J Comput Chem* 2005;**26**:1701–18. https://doi.org/10.1002/jcc.20291.
66. Lindorff-Larsen K, Piana S, Palmo K. *et al.* Improved side-chain torsion potentials for the Amber ff99SB protein force field. *Proteins* 2010;**78**:1950–8. https://doi.org/10.1002/prot.22711.
67. Hornak V, Abel R, Okur A. *et al.* Comparison of multiple Amber force fields and development of improved protein backbone parameters. *Proteins* 2006;**65**:712–25. https://doi.org/10.1002/prot.21123.
68. Darden T, York D, Pedersen L. Particle mesh Ewald: AnN·log(N) method for Ewald sums in large systems. *J Chem Phys* 1993;**98**:10089–92. https://doi.org/10.1063/1.464397.

Briefings in Bioinformatics, 2025, **26(1)**, bbae715
https://doi.org/10.1093/bib/bbae715
Problem Solving Protocol