Briefings in Bioinformatics, 2025, **26(1)**, bbae644
https://doi.org/10.1093/bib/bbae644
Problem Solving Protocol

TG-CDDPM: text-guided antimicrobial peptides generation based on conditional denoising diffusion probabilistic model

Junhang Cao[1,‡], Jun Zhang[2,‡], Qiyuan Yu[1], Junkai Ji[2], Jianqiang Li[2], Shan He[3], Zexuan Zhu[2,*]

[1]College of Computer Science and Software Engineering, Shenzhen University, Shenzhen 518060, China
[2]National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen 518060, China
[3]School of Computer Science, University of Birmingham, Birmingham B15 2TT, United Kingdom

*Corresponding author. National Engineering Laboratory for Big Data System Computing Technology, Shenzhen University, Shenzhen 518060, China. E-mail: zhuzx@szu.edu.cn

‡Junhang Cao and Jun Zhang co-authors contributed equally to this work.

Abstract

Antimicrobial peptides (AMPs) have emerged as a promising substitution to antibiotics thanks to their boarder range of activities, less likelihood of drug resistance, and low toxicity. Traditional biochemical methods for AMP discovery are costly and inefficient. Deep generative models, including the long-short term memory model, variational autoencoder model, and generative adversarial model, have been widely introduced to expedite AMP discovery. However, these models tend to suffer from the lack of diversity in generating AMPs. The denoising diffusion probabilistic model serves as a good candidate for solving this issue. We proposed a three-stage Text-Guided Conditional Denoising Diffusion Probabilistic Model (TG-CDDPM) to generate novel and homologous AMPs. In the first two stages, contrastive learning and inferring models are crafted to create better conditions for guiding AMP generation, respectively. In the last stage, a pre-trained conditional denoising diffusion probabilistic model is leveraged to enrich the peptide knowledge and fine-tuned to learn feature representation in downstream. TG-CDDPM was compared to the state-of-the-art generative models for AMP generation, and it demonstrated competitive or better performance with the assistance of text description as supervised information. The membrane penetration capabilities of the identified candidate AMPs by TG-CDDPM were also validated through molecular weight dynamics experiments.

Keywords: antimicrobial peptides; diffusion model; text guidance; pre-training; fine-tuning

Introduction

The overuse of antibiotics has led humanity into a crisis of escalating antimicrobial resistance. In 2019, drug-resistant infections were linked to an alarming 4.95 million deaths globally, and this number is expected to reach 10 million by 2050 [1, 2]. Antimicrobial peptides (AMPs), a type of short amino acid sequences normally ranging in length from 5 to 50, are investigated as a viable solution to combat antimicrobial-resistant disease treatment due to their low propensity for inducing drug resistance [3]. Traditional methods for discovering novel and potent AMPs rely on biochemical strategies, which usually are costly and time-consuming.

Deep generative models have been widely used to increase the pace of AMP discovery by enhancing overall efficiency and reducing resource consumption during the exploration process. For example, the Long-Short Term Memory (LSTM) model [4] has many applications within peptide generation. Porto *et al.* [5] designed a unique model, namely Joker, to generate peptides based on the principles of card games and sliding windows. Müller *et al.* [6] trained an LSTM model with α-helix AMPs to facilitate the design of new peptides via sampling in the feature space. Nagarajan *et al.* [7] advanced the utilization of the LSTM model to generate active AMPs and synthesized an array of 10 potent peptides, their minimum inhibition concentration (MIC) predicted by a trained bi-LSTM model. Grisoni *et al.* [8] refined the approach by pre-training an LSTM model with a series positive α-helix cationic amphipathic peptides and fine-tuning with 26 recognized active anticancer peptides for de novo design. Wu *et al.* [9] utilized a Transformer framework with an advanced attention mechanism for signal peptide generation. Wang *et al.* [10] adopted two LSTM models to sample AMPs with specified lengths of 15 and 20 respectively. Schissel *et al.* [11] built a closed-loop framework, including an LSTM generator, a predictor, and a genetic algorithm optimizer, towards the generation of membrane-permeable peptides. Zhang *et al.* [12] proposed LSTM_Pep, a model pre-trained on a large number of peptides and fine-tuned to design de novo peptides with specific therapeutic potential. Despite the advancements represented by LSTM-based models in peptide generation, this unparalleled generative autoregressive approach, which generates a current token based on the preceding one, results in lower speed while encountering an issue known as gradient vanishing, especially as the complexity of the model is scaled up with an increase in the number of neurons.

Received: July 5, 2024. **Revised:** November 13, 2024. **Accepted:** November 27, 2024

In light of these challenges, the non-autoregressive generative models like variational autoencoder model (VAE) and generative adversarial model (GAN) have emerged as powerful tools in the field of peptide generation due to their ability to harness latent spaces and to generate all the tokens in parallel. For example, Das *et al.* [13] collected unlabeled peptides for training PepcVAE to gain rich latent information and trained a classifier with labeled peptide sequences. The output from the classifier was used as a conditional input to guide the design of AMPs. Similarly, Dean *et al.* [14] exploited the capability of VAE for reconstructing known AMPs, enabling it to sample novel AMPs with targeted physicochemical properties. Das *et al.* [15] leveraged the guidance from the classifier to train a VAE with an informative latent space of molecules. This method facilitated the generation of novel AMPs via rejection sampling, followed by a screening process using a deep-learning classifier to ensure the retention of desired physicochemical properties. Szymczak *et al.* [16] proposed HydrAMP, a conditioning VAE designed to capture class priors and lower-dimensional representation of peptides. This method disentangles antimicrobial conditions from the representation of peptides for diverse tasks in peptide generation, including unconstrained and analogous generation. Tucs *et al.* [17] put forward PepGAN to generate active AMPs, taking probability distribution between covering active peptides and dodging in-active peptides. PepGAN can generate AMPs with specific physicochemical features. Rossetto *et al.* [18] introduced GANDALF, a sophisticated framework that employs two GAN models by integrating active atom information to generate peptides and corresponding structures. This method validated the effectiveness of the generated peptide sequences through docking experiments. Ferrel *et al.* [19] proposed AMP-GAN, a framework that combines a GAN with an encoder. This innovative approach used conditional and latent vectors obtained through the encoder to guide the GAN in generating corresponding active peptides. Building on the success of AMP-GAN, Oort *et al.* [20] updated their model structure and used bidirectional conditional GAN to propose AMP-GAN v2 for AMP design. Li *et al.* [21] extended the application of GAN and proposed the DeepImmuno-GAN model to generate immune gene sequences. Surana *et al.* [22] applied the Leak-GAN model for the task of antiviral peptide generation and proposed PandoraGAN.

The aforementioned advancements highlight the adaptability and potential of VAEs and GANs in contributing significantly to the field of peptide design. However, the diversity of peptides produced by these models encounters certain constraints due to the sampling process derived from fixed probability distribution or mode collapse [23–25]. Denoising Diffusion Probabilistic Model (DDPM) [26] employs a neural network to simulate the nosing and denoising processes within a Markov chain, serving as a good solution to this issue. DDPM reveals its potential by efficiently generating a diverse array of samples from the Gaussian distribution and producing specific samples aligned with targeted requirements, ranging from image generation [27] to protein design [28]. Inspired by the success of text-guided image and short video generation, we propose a three-stage pre-trained framework based on a text-guided conditional DDPM, namely TG-CDDPM, for AMP generation. The first two stages introduce contrastive learning and inferring models to enhance the conditions for guiding AMP generation, respectively. In the third stage, a pre-trained DDPM is used to enrich the peptide knowledge and fine-tuned to learn feature representation in AMP generation. We utilize the textual descriptions of the AMPs as supervised information to control the sampling distribution and generation efficiency. The mode of pre-training and fine-tuning has been acknowledged for its ability to enrich the feature diversity in latent space, thereby elevating the model's overall performance. The effectiveness of TG-CDDPM was demonstrated in the comparison studies with other state-of-the-art AMP generative models and the molecular weight dynamics experiment. The contributions of this study can be summarized as follows:

- A peptide dataset for pre-training was collected from the Uniprot database and subsequently preprocessed through a sliding window technique. A fine-tuning dataset was collected from a few public databases, including AMPs and corresponding textual descriptions. The collected datasets can serve as a valuable benchmark for future work in peptide generation.
- A three-stage framework, known as TG-CDDPM, was proposed to address the complexities in correlating texts with peptides, enhancing the quality of condition vectors, and generating novel and homologous AMPs, respectively. Unlike LSTM, VAE, and GAN-based models, which focus on peptide generation through sampling in latent space without effective guidance, TG-CDDPM incorporates peptide attribute information in a text-guidance manner. To our knowledge, this is the first attempt in this field.
- The effectiveness of TG-CDDPM was rigorously assessed through comparative experiments. Additionally, the applicability of the generated peptides was further validated via molecular dynamics simulation against both Gram-negative and Gram-positive bacterial membranes. These simulations demonstrated the membrane permeability of the generated candidates, underlining the potential of TG-CDDPM in contributing innovative solutions to the field.

Materials and methods

Datasets

To train TG-CDDPM, we constructed a pre-training dataset with simulated peptides based on UniProt(https://www.uniprot.org/) (see Fig. 1(a)) and a fine-tuning dataset containing known AMPs and inactive peptides with the corresponding descriptive texts from several public databases. Among the fine-tuning dataset, 58 481 APMs were collected from DBAASP v3.0(https://dbaasp.org/) (14169), dbAMP v2.0(https://awi.cuhk.edu.cn/dbAMP/) (20147), CAMP(http://camp.bicnirrh.res.in/) (15512), APD3(https://aps.unmc.edu/AP/) (4173), and Dramp v3.0(http://dramp.cpu-bioinfor.org/) (4480). The 58 443 inactive peptides were collected from AMPlify[29] with a length of 50, used as a control for AMPs. The balance between inactive and active peptides can prevent data biases from influencing the model's learning process and ensure the reliability and diversity of generated AMPs. The processes of data collection are illustrated in Fig. 1.

For an objective and rigorous dataset, we filtered out peptides containing non-natural amino acids, specifically 'B', 'J', 'O', 'U', 'X', and 'Z', and removed redundant peptides with sequence similarity of 90% from the collected data to reduce homology bias, using the CD-HIT program [30]. Considering that the full-text description may contain noise that increases the complexity of model reasoning, we deleted excessive words to retain the most informative description, including source, activity, and taxonomy.

We set a maximum generation length of 50 for TG-CDDPM, aligning with the normal active AMP length range from 5 to 50 [31]. Considering a large portion of pre-training data exceeded length by 50, simply truncating peptides at a fixed length would potentially lose crucial biochemical information within amino

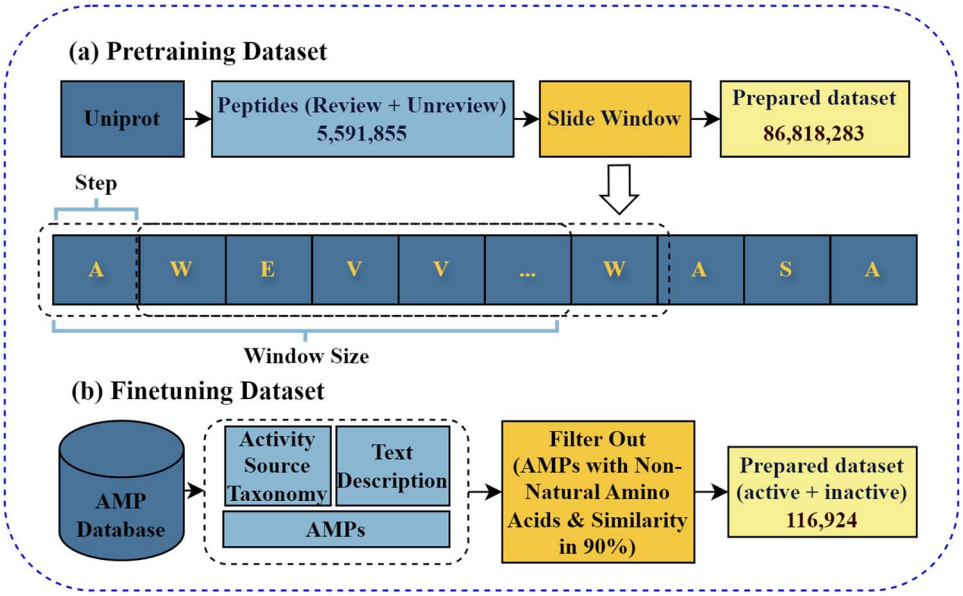


Figure 1. The process of pre-training and fine-tuning datasets collection.

acids. Thus, we leveraged the slide window technique to preserve as much biochemical information as possible from peptides exceeding the length limit. The sliding-window strategy is illustrated in Fig. 1(a). The window size is 50, and the sliding step is 1. Only a few of the AMPs in the fine-tuning dataset exceed the threshold of 50 in length, so we just removed them from the dataset. Finally, we obtained a pre-training dataset with 86 818 283 peptides and a fine-tuning dataset with 116 924 pairs of antimicrobial and inactive peptides, along with their text description as shown in Fig. 1(b). During pre-training and fine-tuning processes, we pad the peptides shorter than 50 at the end with a special symbol, allowing the model to handle variable-length peptides. TG-CDDPM is a Transformer-based diffusion model that can use the attention mechanism to concentrate on the peptide region in a padded sequence. Therefore, the variable length processing steps do not affect the model's performance, especially the peptide generation.

The statistical information of the datasets and existing benchmarks are listed in Table S2 of the Supplementary Materials. To our knowledge, the constructed pre-training dataset is the largest known simulated peptide dataset, and the constructed fine-tuning dataset is the first peptide-text pair dataset for text-guided peptide generation. Compared with existing benchmark datasets, the datasets used in this study extended the length of peptides to 50. Peptides in other datasets are no longer than 32, and some peptides are simply cut during the data collection process, which could lead to information loss. Therefore, the datasets used in this study are more informative and well-suited for potential peptide generation tasks, leveraging pre-training and fine-tuning. Furthermore, the text descriptions could be enhanced with other labels, such as structural information, to advance text-guided peptide generation toward target structures.

Overview of TG-CDDPM

The framework of the proposed TG-CDDPM is depicted in Fig. 2(a). It comprises three modules namely text description encoder, conditional diffusion model, and peptide decoder. The text description encoder translates the peptide description into a latent vector, which is used as a condition to control peptide generation. The conditional diffusion model receives Gaussian noise and the text condition vector to perform inference to yield latent features of AMPs. The peptide decoder generates a peptide sequence according to the latent features yielded by the diffusion model.

The structure of the text description encoder is illustrated in Fig. 2(b). The description text undergoes a transformation into an embedding matrix through the Tokenizer and Embedding Layer. Since the text descriptions of peptides used in this work are mainly derived from scientific materials, we employed SciBert [32] to extract the text features and obtain accurate text representation. SciBert is a sophisticated language model pre-trained with rich online scientific textual materials. Note that other specific language models, such as BioBERT [33] and BLURB [34], are also applicable to this task. Nevertheless, considering the dataset text labels might come from more general areas, SciBert is adopted in this work. To avoid the anisotropy problem of the language model and make the text representation more suitable for the AMP design task, we introduce a trainable adapter subsequent to the SciBERT model to transform the latent features further to adapt peptide generation.

Figure 2(c) depicts the conditional diffusion model, which mainly involves two processes, i.e. adding noise and denoising. In the process of adding noise, the embedding matrix undergoes a forward diffusion transformation, resulting in a matrix of Gaussian noise. We utilize a Transformer model for the denoising process, which comprises multiple blocks with a multi-layer perceptron neural network and multi-head attention mechanisms. The Transformer's excellent learning ability of correlations between different positions enables the diffusion model to capture the complex features of noise in samples. To generate active and novel peptide sequences, the textual condition obtained in the text description encoder is concatenated with a noised embedding matrix, serving as a guide throughout the process.

The peptide decoder consists of a feed-forward neural network with the softmax activation function. It outputs the probability distribution of amino acids at each position of the target peptide based on the latent feature matrix generated by the diffusion model.

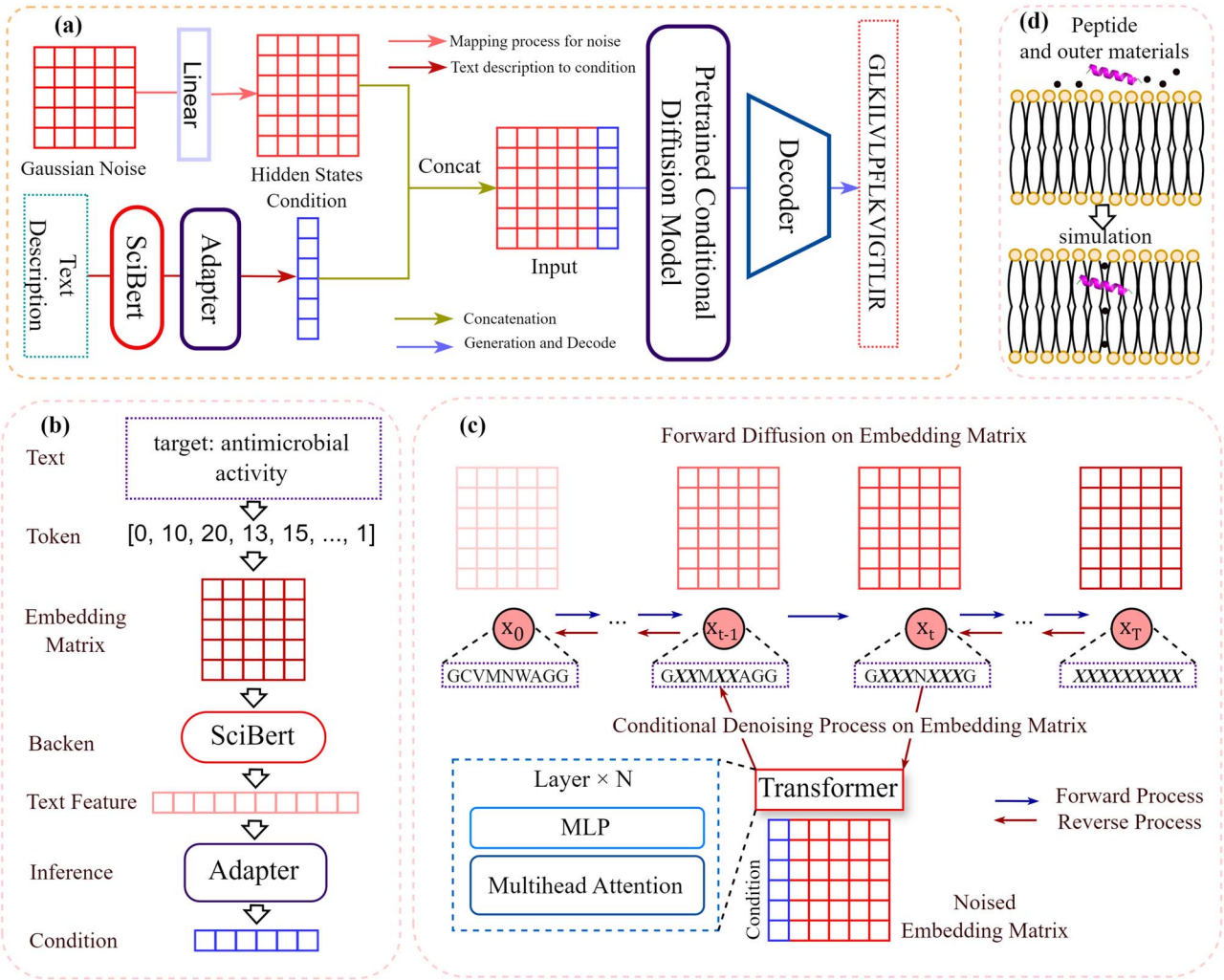


Figure 2. The illustration of the proposed method. (a) The overview of TG-CDDPM framework. (b) The process of the first and second stages in TG-CDDPM. (c) The process of the third stage in TG-CDDPM. (d) The brief illustration of molecular dynamics simulation.

In this study, we designed a three-stage training strategy to minimize potential biases within the computational model, ensuring that the trained model consistently generates reliable peptide candidates. The training of TG-CDDPM was conducted using an NVIDIA RTX GeForce 3050 GPU with 8 GB of memory, completing the process in 72 hours. The key components of TG-CDDPM are detailed as follows.

Text and peptide alignment

The alignment of peptides and text is a key to text-guided peptide generation. Contrastive learning provides an effective way to achieve this goal. In this study, we use CLIP [35], a classical contrastive learning framework, to perform the first-stage pre-training to construct the association between text and peptides. The CLIP framework has been successfully applied to image-text alignment in the field of image generation. It can maximize the similarity between the text and their corresponding peptide while reducing that among mismatched pairs to learn the association between text and peptides effectively. We use the language model SciBERT to encode text and design a peptide encoder based on Transformer neural networks. The amino acids can be regarded as the 'words' of a peptide sequence, so we can define a peptide vocabulary consisting of 20 natural amino acids. As such, natural language processing techniques can be applied to process peptide sequences. The pre-training process of stage one is shown in Fig. 3(a). Let T_{O1} and P_{O1} denote the first pair of text and peptide feature vectors, respectively, T_{O2} and P_{O2} correspond to the second pair. The objective of the training loss function is to narrow the distance between T_{O1} and P_{O1} while maximizing the distance between T_{O1} and P_{O2}. The training hyperparameters of peptide encoder and SciBERT are provided in Table S1 of the Supplementary Materials. The SciBERT remains fixed at this stage. InfoNCE [36] function was used to calculate the loss, which is a classical contrastive learning loss function. It can be formulated by Equation (1):

$$\mathrm{Loss}(x, K) = \mathbb{E}_{p(i|x,K)}\left[-\log \frac{\exp(S(x, k^{+}))}{\sum_{k \epsilon K} \exp(S(x, k))}\right] \quad (1)$$

where $S(\cdot)$ indicates the softmax activation function, x and k^{+} represent the same pair of text and peptide features, respectively, and K denotes the set of all peptide features. The contrastive learning model was trained by the pairs of text and peptides in the fine-tuning dataset. The peptide encoder's weights and biases were updated and transitioned from this stage to the subsequent stages.

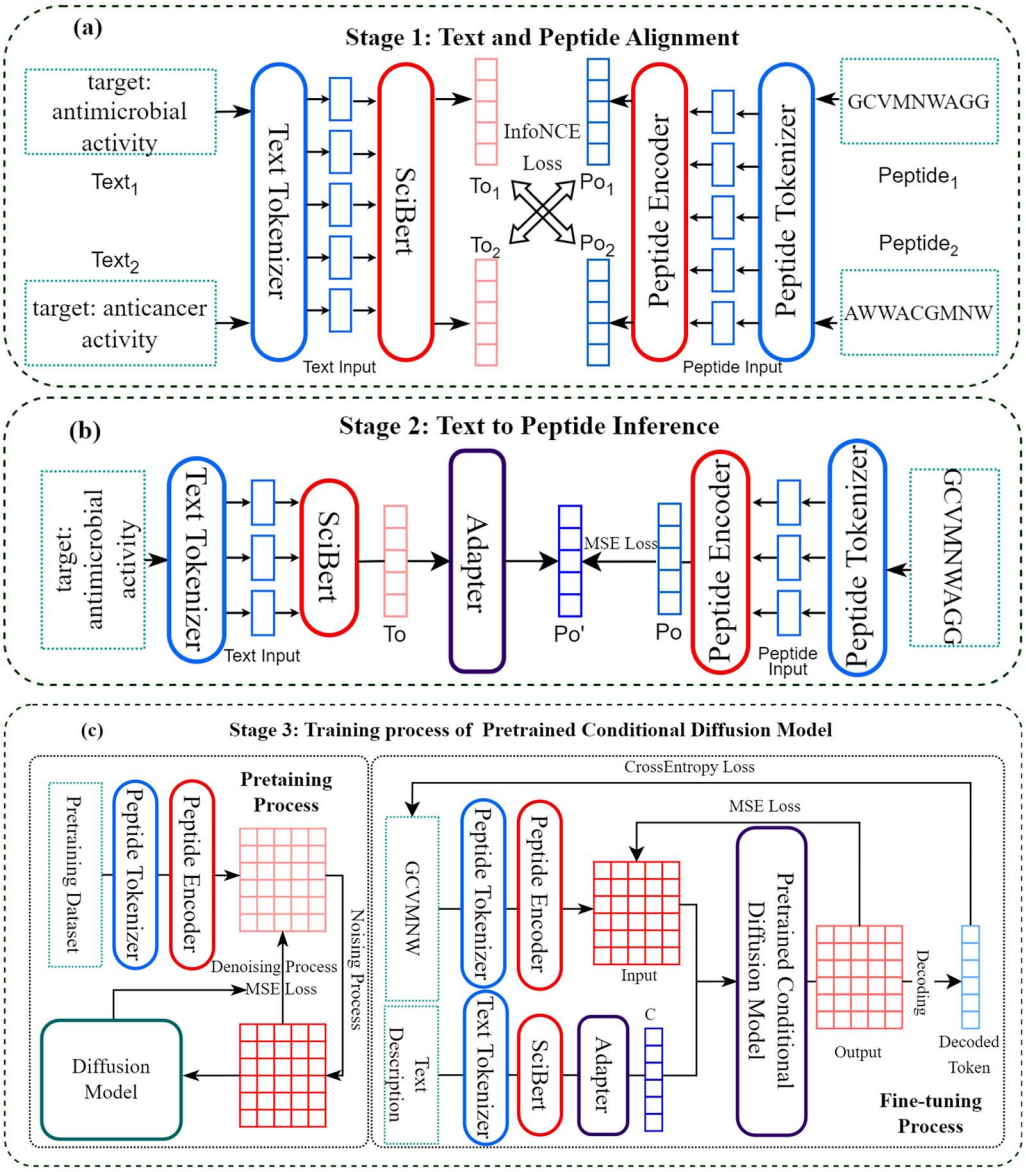


Figure 3. Three training stages of TG-CDDPM. (a) The pre-training process of text and peptide alignment. (b) The training process of text to peptide inference. (c) The pre-training and fine-tuning processes for AMP generation.

Text to peptide inference

Inferring the potential peptides from the text description is a complex problem (many-to-many mapping). To avoid the language model's anisotropy problem and obtain a more accurate condition vector, we introduced a DDPM-based adapter in the text encoder. In the second stage, we pre-trained the adapter to make the language model SciBERT more compatible with our task. During the training phase, we fixed the parameters of the language model SciBERT and the peptide encoder and only trained the adapter using the fine-tuning dataset. The pre-training process is illustrated in Fig. 3(b), where T_O indicates a text feature vector, P_O denotes an amino acid feature vector, and $P_{O'}$ represents a condition vector. The training objective is to close the text condition vector to the corresponding peptides in a certain latent space. In this study, we used the Mean Square Error (MSE) as the loss function to achieve this. It can be formulated by Equation (2):

$$\mathcal{L}_{\text{MSE}} = \frac{\sum_{i=1}^{N} \|O_i - I_i\|^2}{N} \quad (2)$$

where O_i and I_i represent the output and input for the ith sample, respectively. N is the total number of samples in a training batch. The hyperparameters used in this stage are listed in Table S1 of the Supplementary Materials. The trained adapter ensures the derivation of conditions from text features, facilitating text-to-peptide transformation.

Text-guided AMP generation

After the previous two stages, we can obtain a conditional vector to guide the denoising process of the diffusion model and generate specific peptides. To improve the stability and generalization ability of the model, we initially pre-trained a DDPM with the pre-training dataset from UniProt without guidance and then applied the pre-trained model to AMP generation via the transfer learning strategy, ensuring the model's adaptation of text-guided generation task.

The pre-training process is shown in the left panel of Fig. 3(c), where noise is added to the embedding matrix on 500 time steps to generate a Gaussian noise distribution. Then, the DDPM

simulates the Markov chain's denoising process to capture the feature distribution of pre-training peptides and reconstruct the original embedding matrix. MSE serves as the loss function to measure the quality of the reconstructed embedding matrix. The detailed hyperparameters for pre-training are listed in Table S1 of the Supplementary Materials.

The fine-tuning process is shown in the right panel of Fig. 3(c) where the fine-tuning dataset is used to train the DDPM to make it more applicable to APM generation. Unlike the pre-training process, the text conditional vector and Gaussian noise jointly guide the DDPM in reconstructing the peptide embedding matrix. Therefore, the model can produce latent peptide embedding matching the text description after fine-tuning. Like the pre-training process, the MSE is used as the loss function to constrain DDPM and minimize the discrepancies between the reconstructed and original embedding matrices to simulate the denoising process. Restoring the peptide sequence from the latent embedding matrix is also essential in the entire process of peptide generation. In this regard, we use the cross-entropy function to measure the quality of the peptide decoder to restore the peptide sequence. The cross-entropy loss function can be formulated as follows:

$$\mathcal{L}_{\mathrm{CE}} = \frac{-1}{N}\sum_{\mathrm{i}}^{\mathrm{N}} [\mathrm{y_i} \cdot \log(q_\mathrm{i}) + (1 - \mathrm{y_i}) \cdot \log(1 - q_\mathrm{i})] \tag{3}$$

where y_i represents the label of the ith amino acid, equivalent to the Token value of the input sequence, and q_i represents the output probability distribution of amino acid at position i.

Therefore, the fine-tuning phase is a multi-objective optimization process. The complete loss function is shown below:

$$\mathcal{L} = \mathcal{L}_{\mathrm{MSE}} + \mathcal{L}_{\mathrm{CE}} \tag{4}$$

where L_{MSE} is the peptide embedding matrix reconstructing loss and L_{CE} is peptide sequence decoding loss. The hyperparameters of fine-tuning are shown in Table S1 of the Supplementary Materials. This detailed fine-tuning strategy substantially enhances the model's ability to recreate intricate peptide sequences, thereby solidifying the decoding efficacy of our denoising diffusion generative model. Within the peptide generation phase, we do random sampling with Gaussian noise and text guidance that describes specific peptide properties, enabling the generation of a wide variety of AMPs.

Evaluation metrics

To evaluate the proposed TG-CDDPM framework, we introduced an evaluation metric named AMscore based on three AMP predictors, including amPEP [37], CAMP [38], and IPPF-FE [39]. The metric measures the proportion of generated peptides with predicted antimicrobial activity as follows:

$$\mathrm{AMscore} = \frac{N_{\mathrm{am}}}{N} \tag{5}$$

where N_{am} is the number of peptides with predicted antimicrobial activity and N represents the total number of generated peptides. The predicted antimicrobial activity is calculated by amPEP, CAMP, and IPPF-FE. For instance, after synthesizing a batch of peptide sequences, we can use a predictor to calculate the likelihood of a sequence being an AMP. If the predicted score of a peptide meets the requirement of the threshold value (i.e. 0.5) [40, 41], the peptide is considered to have antimicrobial activity. Additionally, molecular dynamics simulation was also used to evaluate the proposed method. We used mass density [42] to measure molecular dynamics simulation results.

Table 1. The results of comparison TG-CDDPM with three state-of-the-art methods

	amPEP	CAMP r3	IPPF-FE
PepcVAE	0.583±0.0028***	0.735±0.0027***	0.388±0.0011***
- w/o guidance	0.580±0.0023***	0.533±0.0031***	0.414±0.0025***
ampGAN v2	0.834±0.0022***	0.873±0.0009***	0.819±0.0014 -
HydrAMP	0.895±0.0006***	0.854±0.0023***	0.798±0.0021*
- w/o guidance	0.835±0.0016***	0.751±0.0039***	0.767±0.0019***
TG-CDDPM	**0.964**±0.0003	**0.977**±0.0002	**0.833**±0.0005
- w/o guidance	0.866±0.0013***	0.843±0.0019***	0.774±0.0007***

Results

Comparison with state-of-the-art methods

TG-CDDPM was compared with three representative state-of-the-art deep-learning-based AMP generation methods, namely PepcVAE [13], ampGAN v2 [20], and HydrAMP [16]. It costs almost 8 days in batch size 20 for TG-CDDPM to generate millions of peptides. In a single comparison experiment, each method was configured to generate 100 peptides. We used the text: 'This is a peptide: target antibacterial, anticancer antiviral' in TG-CDDPM to guide the generation of 100 peptides. PepcVAE employs minimum inhibitory concentration (MIC) probabilities as guidance to generate 100 peptides, and ampGAN v2 uses a series of labels (including sequence length, MIC, target microbes, and target mechanisms) to instruct the model to generate 100 peptides. HydrAMP requires existing peptides as templates for guidance. The quality of selected reference peptides impacts HydrAMP's generations. For an objective comparison, we randomly sampled 100 collected test peptides as templates to guide HydrAMP for the generation. To explore the role of guidance information, we also evaluated the performance of different models without guidance, except ampGAN v2, since it does not provide an unguided option.

The experiment was repeated 10 times, and the average and standard deviation of the AMscores were calculated. A Student's t-test [43] was conducted to determine the significance of the differences between TG-CDDPM and other models. The significance levels of t-test are denoted as: -: $p>0.05$, *: $p\leq0.05$, **: $p\leq0.01$, and ***: $p\leq0.001$. The results of different methods are listed in Table 1. TG-CDDPM outperforms the counterpart competing methods w/o text guidance in terms of AMscore. It is worth noting that the performance of all methods decreased after the guidance module was removed. This indicates that the guidance information is valuable for generating AMPs. In addition, compared to the other two methods with guidance information, i.e. PepcVAE and HydrAMP, TG-CDDPM showed more significant improvement with text guidance, demonstrating better capability of cooperating with diverse input features. The confidence intervals for TG-CDDPM with guidance, in amPEP, CAMP r3, and IPPF-FE are $[0.950, 0.978]$, $[0.966, 0.988]$, and $[0.816, 0.849]$, respectively.

To validate the antimicrobial properties of the generated peptides, we analyzed three critical physicochemical properties related to the antimicrobial activity of 100 generated peptides, including charge, GRAVY hydrophobicity, and TM tend moment. These properties were calculated by a Python package, Modlamp [44]. To objectively compare with other methods, we also

Table 2. Top 10 generated candidate peptides

Candidates	hits	e-value	Secondary structure
IVASIKARLGKLI	1	9.0279	α-helix
FLGPALLVAHGLVKGRC	3	8.9362	α-helix
GIKPLLNKKLSGL	1	8.8821	α-helix
GLKILVLPFLKVIGTLIR	7	7.9600	hybrid-helix
WKGKLAARLALLKLL	1	7.5218	α-helix
GFLKLAVRRKKRVNALC	2	6.3437	α-helix
GWLKLLKKAKKVIIGVM	2	5.3093	α-helix
GWLKVLKPRLKLI	19	5.2798	3–10-helix
VVPLLFGGLGKKL	52	4.0595	coli and turn
KILGLFKLGKLVVAK	1	3.7602	α-helix
GVKKILVAAKKL	127	2.9669	α-helix

conducted a significance analysis by the Student's t-test. The results are shown in Fig. S1 of the Supplementary Materials. TG-CDDPM generated peptides with an average positive charge of +6, outperforming other methods. The same results were observed in terms of the GRAVY hydrophobicity. While TG-CDDPM achieved the third-best TM tend moment, it is comparable to those produced by HydrAMP and ampGAN v2. These results indicate that the AMPs generated by TG-CDDPM are reliable compared with existing state-of-the-art computational methods.

Molecular dynamics simulation

To estimate the activeness of the generated peptides by TG-CDDPM, we validated the propensity for cell membranes of the generated peptides via a molecular dynamics simulation experiment targeting the interaction of the peptides with both Gram-negative and Gram-positive bacterial membranes. We used TG-CDDPM to generate 100 peptide sequences for evaluation. The generated peptides could be homologous with known AMPs in the fine-tuning dataset, the activeness of which has been verified. Our interest is in novel peptides that are distinguished from the known ones. New peptides might better serve as innovative drug candidates. In this regard, we utilized the BLAST tool [45] to filter the generated peptides against the known AMPs in the fine-tuning dataset, ensuring the generated peptides cover diverse antimicrobial properties of active peptides in terms of homology. The e-value of BLAST measures the homogeneity between the generated and the known AMPs. A larger e-value indicates better singularity of the corresponding generated peptide. The number of hits indicates the number of known AMPs matched to a candidate peptide by BLAST. The top 10 generated peptides in terms of e-value are tabulated in Table 2.

We used GROMACS [46] to perform 100ns molecular dynamics simulations between the generated peptides and the Gram-negative and Gram-positive bacterial membranes. Online utility CHARMM-GUI [47] was used for constructing Gram-negative and Gram-positive bacterial membranes and the corresponding force fields. We set the ratio of the molecules forming the Gram-negative bacterial membrane as POPE/POPG/TOCL to 7:2:1, and the ratio of the Gram-positive bacterial membrane as POPG/POPE/DOTAP/TOCL to 6:1.5:1.5:1 [48]. The conditions for molecular dynamics simulation, including force fields, temperature, and pressure, were automatically generated without manual setting, and the corresponding files and codes are available in the GitHub repository. The mass density [42] metric was then applied in evaluating the actual interaction between the peptides and membranes and identifying which conformations most significantly influence membrane penetration tendency.

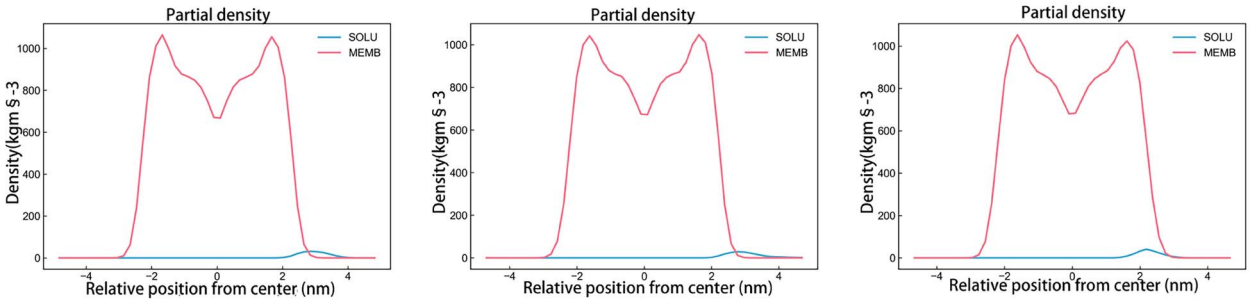
Based on the three-dimensional conformation prediction of AlphaFold 2 [49], a predictor available on the online platform, the pTMscore was calculated to validate the accuracy of structures. The peptides with the highest score can be mainly classified into three types, i.e. α-helix structure, 3–10 helix structure, and hybrid structures comprising both. The mass density distributions of three representative peptides, each for a secondary structure (highlighted in Table 2), over Gram-negative bacterial membranes are shown in Fig. 4. Among the representative peptides, the peptide 'GIKPLLNKKLSGL' bears an exclusive α-helix structure, and 'GLKILVLPFLKVIGTLIR' bears a hybrid helix structure of both α-helix and 3–10 helix, while 'GWLKVLKPRLKLI' has a concise 3–10 helix. In Fig. 4, the horizontal and vertical axes symbolize the membrane's center position and the mass density, respectively, with the red and blue curves denoting the positions of membrane and peptide. From time 0 to 100 ns, all three peptides manifested progressive interaction with the membrane. 'GIKPLLNKKLSGL' and 'GLKILVLPFLKVIGTLIR' show stronger membrane penetration capability and are embedded deeper into the membrane, as shown in Figs 4(a) and 4(b). 'GWLKVLKPRLKLI' exhibited a relatively weaker interaction, as shown in Fig. 4(c). The visualization of the simulation results by VMD fool [50] is shown in Fig. 5, which provides a more intuitive view of the candidate peptides with secondary structure on the membrane permeability of the Gram-negative bacterial membranes. Similar results were observed in simulations involving Gram-positive bacterial membranes, with mass density and visualization results illustrated in Figs S2 and S3 of the Supplementary Materials, respectively. The results demonstrated the membrane permeability tendency of the generated peptides in Gram-negative and Gram-positive bacterial membranes and the secondary structure can impact AMP membrane permeability reaction [51], i.e. the α-helix structure may be a vital element that affects the degree of membrane permeability in AMPs.

Minimum inhibition concentration analysis

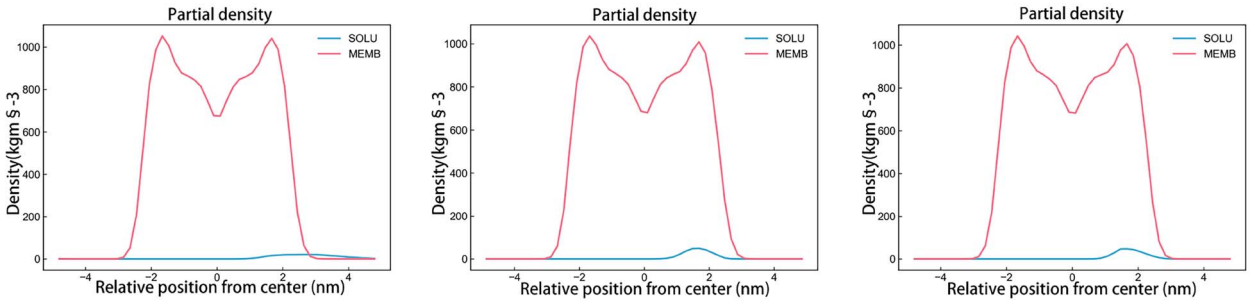
MIC is a classical indicator used to measure the antibacterial activity of antibiotics. However, conventional laboratory experiments of MIC are time-consuming. Computation-based methods [52, 53] have served as an important supplement to experimental techniques for analyzing antimicrobial activity. To further validate the performance of the generated AMPs, we used AMPActiPred [53], a computational model designed to predict the antibacterial activity of AMPs against various microbes, to estimate the MIC of the generated candidate AMPs. We first validated the reliability of AMPActiPred on known natural AMPs and inactive peptides, as shown in Fig. S4(a) of the Supplementary Materials. AMPActiPred was then used to predict the MIC values of the top 10 generated AMP candidates as shown in Table 3. The results show that the generated peptides demonstrate activity against various bacteria. Notably, nearly all tested peptides were predicted to be active against *S.aureus*.. More details of the predictions are also available in Fig. S4(b) of the Supplementary Materials.

Ablation experiment

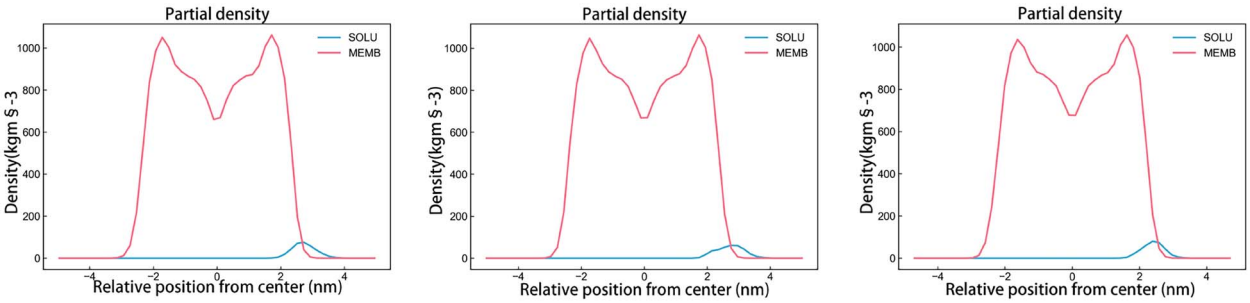
To investigate the roles of different modules in the proposed method, we conducted an ablation experiment on TG-CDDPM. For this purpose, we set up three control variants, i.e. a model



(a) GIKPLLNKKLSGL



(b) GLKILVLPFLKVIGTLIR



(c) GWLKVLKPRLKLI

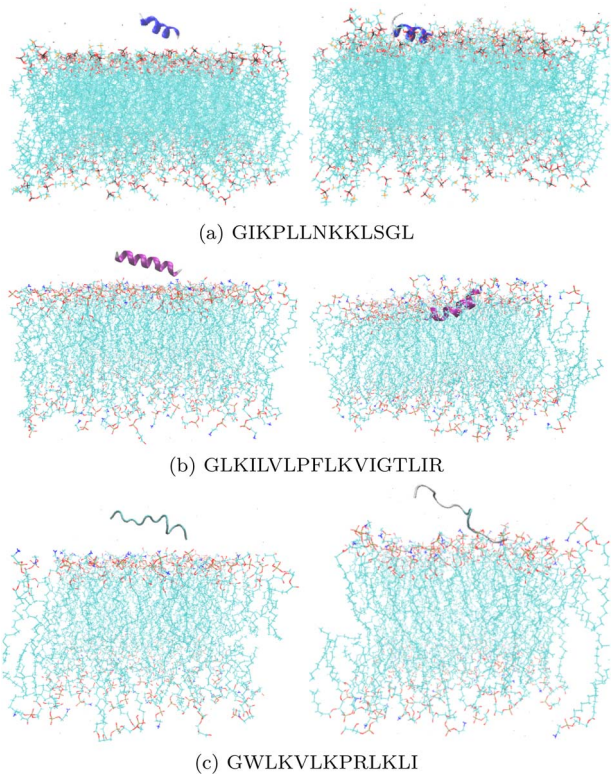
Figure 4. Mass density for generated candidate peptides targeting Gram-negative bacterial membrane, across three-time intervals: 0–25, 25–75, and 75–100 ns, from left to right.

TG-CDDPM[1] without pre-training, a model TG-CDDPM[2] without the second-stage training, and a model TG-CDDPM[3] without second-stage training or pre-training. Each model used the text 'This is a peptide: target antibacterial, anticancer antiviral' to guide TG-CDDPM to generate 100 peptides. We repeated the evaluation experiment 10 times and calculated the average and standard deviation of the results. The AMscores of different models were calculated by three AMP predictors. We also conducted the statistical tests via a Student's t-test to determine the significance of the differences between TG-CDDPM and other models. The results are reported in Table 4. As can be seen in the results, skipping both the pretraining and the second stage of training, TG-CDDPM[3] suffered from a significant performance deterioration. With either one of them, TG-CDDPM[1] or TG-CDDPM[2] can attain improvement from TG-CDDPM[3], which indicates the effectiveness of these two modules. TG-CDDPM[1] is better than TG-CDDPM[2], suggesting that the second-stage training might contribute more to the performance of TG-CDDPM. It is evident that the complete TG-CDDPM achieves the best overall performance in terms of the three metrics by conducting both the second-stage training and pre-training together.

We also investigated the influence of different fine-tuning strategies and input text guidance on the method. We compared the quality of generated peptides using supervised and unsupervised fine-tuning, respectively. The results are reported in Table S3 of the Supplementary Materials. TG-CDDPM achieved higher success rates for active AMP generation with text guidance when fine-tuned with supervised information, enhancing the specificity of generated peptides. Results of the ablation experiment on input text are summarized in Table S4 of the Supplementary Materials,

Table 3. Minimum inhibitory concentration of the top 10 generated candidate peptides against Gram-positive and Gram-negative bacteria based on AMPActiPred. The '–' indicates that the peptide was predicted as inactive against the corresponding microbe

Candidate	Gram-positive Activity (log MIC, unit: uM)			Gram-negative Activity (log MIC, unit: uM)		
	B.subtilis	E.faecalis	S.aureus	K.pneumoniae	P.aeruginosa	A.baumannii
IVASIKARLGKLI	–	1.4926	1.3951	–	1.4812	–
FLGPALLVAHGLVKGRC	–	–	1.2483	1.5175	–	1.0574
GIKPLLNKKLSGL	–	–	1.5385	1.4203	–	–
GLKILVLPFLKVIGTLIR	–	–	0.8400	–	–	–
WKGKLAARLALLKLL	–	–	0.7632	–	–	–
GFLKLAVRRKKRVNALC	–	–	1.1731	–	1.0484	1.0628
GWLKLLKKAKKVIIGVM	–	–	0.8482	–	1.0372	–
GWLKVLKPRLKLI	0.8651	–	–	1.2133	1.2810	–
VVPLLFGGLGKKL	–	–	1.1196	–	–	–
KILGLFKLGKLVVAK	–	–	1.0844	–	–	–
GVKKILVAAKKL	–	–	1.3518	1.1165	1.5516	–



(a) GIKPLLNKKLSGL

(b) GLKILVLPFLKVIGTLIR

(c) GWLKVLKPRLKLI

Figure 5. VMD visualization for generated candidate peptides targeting Gram-negative bacterial membrane, across time intervals: 0–50 and 50–100 ns, from left to right.

Table 4. The results of ablation study for TG-CDDPM

	amPEP	CAMP r3	IPPF-FE
TG-CDDPM	**0.964**±0.0003	**0.977**±0.0002	0.833±0.0005
TG-CDDPM[1]	0.916±0.0006***	0.960±0.0001*	**0.869**±0.0011*
TG-CDDPM[2]	0.901±0.0007***	0.949±0.0003**	0.783±0.0021**
TG-CDDPM[3]	0.757±0.0023***	0.875±0.0035***	0.628±0.0099***

TG-CDDPM[1]: TG-CDDPM w/o pre-training. TG-CDDPM[2]: TG-CDDPM w/o the 2nd stage training. TG-CDDPM[3]: TG-CDDPM w/o pre-training or the 2nd stage training.

which show that the content and length of the guide text affect the generation of AMPs, and richer and more accurate textual descriptions help TG-CDDPM generate AMPs.

Discussions

While TG-CDDPM demonstrated promising results in generating novel AMPs, several limitations exist that warrant further investigation. First, the length of pre-training peptides is limited to 50 after pre-processing with a sliding window approach. This process inevitably removes some biochemical information, resulting in an incomplete representation of the pre-trained model. Second, the fine-tuning dataset lacks detailed descriptions of inactive peptides. To address this problem, we labeled all of them simply as 'target: inactive peptide', which diminishes the diversity of negative descriptive information and affects the generation of inactive peptides. Besides, TG-CDDPM is constrained by a peptide length ≤ 50 for generation, which limits its ability to generate active peptides longer than 50 amino acids. Last, TG-CDDPM relies solely on text description to guide peptide generation. While the function of peptides largely depends on their spatial structure. Therefore, TG-CDDPM may be extended to a model that can generate peptides of any length. TG-CDDPM was trained with general text description including target, source, and taxonomy instead of specific supervised information showing the antimicrobial activities of peptides, such as charge, hydrophobicity, and structures. Incorporating these specific descriptions into the training datasets could address the problem of missing textual description information for inactive peptides and enhance active peptide design.

Peptides are considered intermediaries between amino acid and protein sequences, capable of transitioning to proteins through the folding of secondary structures. Given the similarities between peptides and proteins, TG-CDDPM may also be adapted for protein sequence design with appropriate text descriptions. Moreover, TG-CDDPM employs a self-attention mechanism and a positional encoding module, which could enable it to focus on critical positional or structural information for designing protein secondary structures, provided that structural data is available. Through the application of TG-CDDPM, we can explore a wide landscape for peptide and protein design.

Conclusion

In this study, we propose a pre-trained generative diffusion model with text guidance, TG-CDDPM, for AMP generation aimed at reducing resource consumption in biochemical methodologies. Comparative analyses demonstrate that TG-CDDPM outperforms state-of-the-art methods in AMP generation. Additionally, molecular dynamics simulations indicate that the generated AMPs are

capable of membrane penetration, underscoring TG-CDDPM's potential as a tool for discovering effective peptide-based therapeutics. TG-CDDPM facilitates the exploration and identification of active AMPs prior to biochemical and clinical testing, thereby accelerating the AMP discovery process.

Despite its efficacy, TG-CDDPM has a limitation that it currently cannot generate AMPs with specific secondary structures, which are crucial for AMP activity. Future improvements are deserved to enhance TG-CDDPM to generate AMPs with targeted structures through more rich guidance.

Key Points

- This study constructs a pre-training dataset and a fine-tuning dataset for antimicrobial peptide generation, which can serve as a valuable benchmark for future works.
- This study proposes a multi-stage framework for antimicrobial peptide generation based on text description that outperforms three state-of-the-art methods.
- The molecular dynamics simulation demonstrated the membrane permeability of the generated candidates by the proposed method, showing its potential applicability in drug design.

Funding

This work is supported by the National Key Research and Development Program of China (2022YFF1202100), the National Natural Science Foundation of China (62471310, 62302311, and 62476177), and the Guangdong Basic and Applied Basic Research Foundation (2024A1515011681).

Data and Code Availability

All the datasets used in this study and the source code of the proposed method are available at https://github.com/JunhangCao/TG-CDDPM.

References

1. Murray CJL, Ikuta KS, Sharara F. *et al*. Global burden of bacterial antimicrobial resistance in 2019: A systematic analysis. *The Lancet* 2022; **399**:629–55. https://doi.org/10.1016/S0140-6736(21)02724-0.
2. O'Neill J. *Tackling Drug-Resistant Infections Globally: Final Report and Recommendations*, London, The Review on Antimicrobial Resistance, 2016.
3. Huang J, Yanchao X, Xue Y. *et al*. Identification of potent antimicrobial peptides via a machine-learning pipeline that mines the entire space of peptide sequences. *Nat Biomed Eng* 2023; **7**:797–810. https://doi.org/10.1038/s41551-022-00991-2.
4. Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput* 1997; **9**:1735–80. https://doi.org/10.1162/neco.1997.9.8.1735.
5. Porto WF, Fensterseifer ICM, Ribeiro SM. *et al*. Joker: An algorithm to insert patterns into sequences for designing antimicrobial peptides. 2018; **1862**:2043–52. https://doi.org/10.1016/j.bbagen.2018.06.011.
6. Muller AT, Hiss JA, Schneider G. Recurrent neural network model for constructive peptide design. *J Chem Inf Model* 2018; **58**:472–9. https://doi.org/10.1021/acs.jcim.7b00414.
7. Nagarajan D, Nagarajan T, Roy N. *et al*. Computational antimicrobial peptide design and evaluation against multidrug-resistant clinical isolates of bacteria. *J Biol Chem* 2018; **293**:3492–509. https://doi.org/10.1074/jbc.M117.805499.
8. Grisoni F, Neuhaus CS, Gabernet G. *et al*. Designing anticancer peptides by constructive machine learning. *ChemMedChem* 2018; **13**:1300–2. https://doi.org/10.1002/cmdc.201800204.
9. Zachary W, Yang KK, Liszka MJ. *et al*. Signal peptides generated by attention-based neural networks. *ACS Synth Biol* 2020; **9**:2154–61. https://doi.org/10.1021/acssynbio.0c00219.
10. Wang C, Garlick S, Zloh M. Deep learning for novel antimicrobial peptide design. *Biomolecules* 2021; **11**:471. https://doi.org/10.3390/biom11030471.
11. Schissel CK, Mohapatra S, Wolfe JM. *et al*. Deep learning to design nuclear-targeting abiotic miniproteins. *Nat Chem* 2021; **13**:992–1000. https://doi.org/10.1038/s41557-021-00766-3.
12. Zhang H, Saravanan KM, Wei Y. *et al*. Deep learning-based bioactive therapeutic peptide generation and screening. *J Chem Inf Model* 2023; **63**:835–45. https://doi.org/10.1021/acs.jcim.2c01485.
13. Das P, Wadhawan K, Chang O. *et al*. PepCVAE: semi-supervised targeted design of antimicrobial peptide sequences. ArXiv Preprint ArXiv:1810.07743. 2018. https://arxiv.org/abs/1810.07743.
14. Dean SN, Walper SA. Variational autoencoder for generation of antimicrobial peptides. *ACS Omega* 2020; **5**:20746–54. https://doi.org/10.1021/acsomega.0c00442.
15. 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.
16. 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.
17. 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.
18. Rossetto A, Zhou W. Gandalf: peptide generation for drug design using sequential and structural generative adversarial networks. In: *Proceedings of the 11th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*. New York, United States, ACM, pp. 1–10, 2020.
19. Ferrell JB, Remington JM, Van Oort CM. *et al*. A generative approach toward precision antimicrobial peptide design. *BioRxiv*, 2020:2020–10. https://doi.org/10.1101/2020.10.02.324087.
20. Van Oort CM, Ferrell JB, Remington JM. *et al*. AMPGAN v2: machine learning-guided design of antimicrobial peptides. *J Chem Inf Model* 2021; **61**:2198–207. https://doi.org/10.1021/acs.jcim.0c01441.
21. 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.
22. Surana S, Arora P, Singh D. *et al*. PandoraGAN: generating antiviral peptides using generative adversarial network. *SN Comput Sci* 2023; **4**:607. https://doi.org/10.1007/s42979-023-02203-3.
23. Qiu Y, O'Connor MS, Xue M. *et al*. An efficient path classification algorithm based on variational autoencoder to identify metastable path channels for complex conformational changes. *J Chem Theory Comput* 2023; **19**:4728–42. https://doi.org/10.1021/acs.jctc.3c00318.
24. Jiliang X, Chungui X, Cao R. *et al*. Generative adversarial network-based data augmentation method for anti-coronavirus

peptides prediction. In: *International Conference on Intelligent Computing*, pp. 67–76. Singapore, Springer, 2023. https://doi.org/10.1007/978-981-99-4749-2_6.
25. Lin T-T, Yang L-Y, Lin C-Y. *et al*. Intelligent de novo design of novel antimicrobial peptides against antibiotic-resistant bacteria strains. *Int J Mol Sci* 2023; **24**:6788. https://doi.org/10.3390/ijms24076788.
26. Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models. *Adv Neural Inf Process Syst* 2020; **33**:6840–51.
27. Esser P, Kulal S, Blattmann A. *et al*. Scaling rectified flow transformers for high-resolution image synthesis. In: *Forty-first International Conference on Machine Learning*. Vienna, Austria, ICML, 2024.
28. Watson JL, Juergens D, Bennett NR. *et al*. De novo design of protein structure and function with rfdiffusion. *Nature* 2023; **620**:1089–100. https://doi.org/10.1038/s41586-023-06415-8.
29. 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**:77. https://doi.org/10.1186/s12864-022-08310-4.
30. Li W, Godzik A. CD-HIT: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 2006; **22**:1658–9. https://doi.org/10.1093/bioinformatics/btl158.
31. Wan F, de la Fuente-Nunez C. Mining for antimicrobial peptides in sequence space. *Nat Biomed Eng* 2023; **7**:707–8. https://doi.org/10.1038/s41551-023-01027-z.
32. Beltagy I, Lo K, Cohan A. SciBERT: a pretrained language model for scientific text. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China, Association for Computational Linguistics, 2019, 3615–20.
33. Lee J, Yoon W, Kim S. *et al*. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics* 2020; **36**:1234–40. https://doi.org/10.1093/bioinformatics/btz682.
34. Yu G, Tinn R, Cheng H. *et al*. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans Comput Healthc* 2021; **3**:1–23. https://doi.org/10.1145/3458754.
35. Radford A, Kim JW, Hallacy C. *et al*. Learning transferable visual models from natural language supervision. In: *Proceedings of the 38th International Conference on Machine Learning, volume 139 of Proceedings of Machine Learning Research*. Meila M, Zhang T (eds.), pp. 8748–63. Virtual Event, PMLR, 2021.
36. van den Oord A, Li Y, Vinyals O. Representation learning with contrastive predictive coding. ArXiv Preprint ArXiv:1807.03748. 2018. https://arxiv.org/abs/1807.03748.
37. Bhadra P, Yan J, Li J. *et al*. AmPEP: sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and Random Forest. *Sci Rep* 2018; **8**:1697. https://doi.org/10.1038/s41598-018-19752-w.
38. Gawde U, Chakraborty S, Waghu FH. *et al*. CAMPR4: a database of natural and synthetic antimicrobial peptides. *Nucleic Acids Res* 2023; **51**:D377–83. https://doi.org/10.1093/nar/gkac933.
39. Han Y, Luo X. IPPF-FE: an integrated peptide and protein function prediction framework based on fused features and ensemble models. *Brief Bioinform* 2023; **24**:bbac476.
40. 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**:bbab439. https://doi.org/10.1093/bib/bbab439.
41. 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.
42. Zhao L, Cao Z, Bian Y. *et al*. Molecular dynamics simulations of human antimicrobial peptide ll-37 in model POPC and POPG lipid bilayers. *Int J Mol Sci* 2018; **19**:1186. https://doi.org/10.3390/ijms19041186.
43. Student. The probable error of a mean. *Biometrika* 1908; **6**:1–25.
44. Müller AT, Gabernet G, Hiss JA. *et al*. modlAMP: Python for antimicrobial peptides. *Bioinformatics* 2017; **33**:2753–5. https://doi.org/10.1093/bioinformatics/btx285.
45. Ye J, McGinnis S, Madden TL. Blast: improvements for better sequence analysis. *Nucleic Acids Res* 2006; **34**:W6–9. https://doi.org/10.1093/nar/gkl164.
46. 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.
47. Jo S, Kim T, Iyer VG. *et al*. CHARMM-GUI: a web-based graphical user interface for charmm. *J Comput Chem* 2008; **29**:1859–65. https://doi.org/10.1002/jcc.20945.
48. Allsopp R, Pavlova A, Cline T. *et al*. Antimicrobial peptide mechanism studied by scattering-guided molecular dynamics simulation. *J Phys Chem B* 2022; **126**:6922–35. https://doi.org/10.1021/acs.jpcb.2c03193.
49. Jumper J, Evans R, Pritzel A. *et al*. Highly accurate protein structure prediction with alphafold. *Nature* 2021; **596**:583–9. https://doi.org/10.1038/s41586-021-03819-2.
50. Humphrey W, Dalke A, Schulten K. VMD: visual molecular dynamics. *J Mol Graph* 1996; **14**:33–8. https://doi.org/10.1016/0263-7855(96)00018-5.
51. Chen N, Jiang C. Antimicrobial peptides: structure, mechanism, and modification. *Eur J Med Chem* 2023; **255**:115377. https://doi.org/10.1016/j.ejmech.2023.115377.
52. Melo MCR, Maasch JRMA, de la Fuente-Nunez C. Accelerating antibiotic discovery through artificial intelligence. *Commun Biol* 2021; **4**:1050. https://doi.org/10.1038/s42003-021-02586-0.
53. Yao L, Guan J, Xie P. *et al*. AMPActiPred: a three-stage framework for predicting antibacterial peptides and activity levels with deep forest. *Protein Sci* 2024; **33**:e5006. https://doi.org/10.1002/pro.5006.

Briefings in Bioinformatics, 2025, **26(1)**, bbae644
https://doi.org/10.1093/bib/bbae644
Problem Solving Protocol