nature materials

Article

https://doi.org/10.1038/s41563-025-02164-3

De novo design of self-assembling peptides with antimicrobial activity guided by deep learning

Received: 11 April 2024

Accepted: 30 January 2025

Published online: 14 March 2025

Check for updates

Huayang Liu[1,2,6], Zilin Song[3,4,5,6], Yu Zhang[1], Bihan Wu[1], Dinghao Chen[1], Ziao Zhou[1], Hongyue Zhang[1], Sangshuang Li[1], Xinping Feng[3,5], Jing Huang[3,4,5] ✉ & Huaimin Wang[1,2] ✉

Bioinspired materials based on self-assembling peptides are promising for tackling various challenges in biomedical engineering. While contemporary data-driven approaches have led to the discovery of self-assembling peptides with various structures and properties, predicting the functionalities of these materials is still challenging. Here we describe the deep learning-guided de novo design of antimicrobial materials based on self-assembling peptides targeting bacterial membranes to address the emerging problem of bacterial drug resistance. Our approach integrates non-natural amino acids for enhanced peptide self-assembly and effectively predicts the functional activity of the self-assembling peptide materials with minimal experimental annotation. The designed self-assembling peptide leader displays excellent in vivo therapeutic efficacy against intestinal bacterial infection in mice. Moreover, it exhibits an enhanced biofilm eradication capability and does not induce acquired drug resistance. Mechanistic studies reveal that the designed peptide can self-assemble on bacterial membranes to form nanofibrous structures for killing multidrug-resistant bacteria. This work thus provides a strategy to discover functional peptide materials by customized design.

Self-assembling functional peptide (SAFP) materials have been rationally designed and applied across various fields due to their ease of synthesis and functionalization[1–6]. For example, Silva et al. designed an amphiphilic peptide by incorporating the binding subsequence (IKVAV) as a self-assembly motif that can form nanofibrous hydrogels[7] to promote neuron differentiation. Yolamanova et al. reported an artificial nanofibrous dodecapeptide that facilitates retroviral gene transfer more efficiently than naturally occurring semen-derived enhancers of viral infection fibrils[8,9]. It has also been reported that incorporating tuning motifs promotes in situ peptide self-assembling that exhibits prominent potential for the fields of cancer therapy and bioimaging[10]. The outstanding function and biocompatibility of SAFPs make them prime candidates for developing biodegradable materials for regenerative medicine and tissue engineering.

However, only a limited number of SAFP materials can be developed via empirical design. Computer-aided approaches have facilitated transformative advances in structural predictions and the design of proteins[11–15]. Large protein language models such as AlphaFold rely

[1]Key Laboratory of Precise Synthesis of Functional Molecules of Zhejiang Province, Department of Chemistry, School of Science, Westlake University, Hangzhou, China. [2]Institute of Natural Sciences, Westlake Institute for Advanced Study, Hangzhou, China. [3]Key Laboratory of Structural Biology of Zhejiang Province, School of Life Sciences, Westlake University, Hangzhou, China. [4]Institute of Biology, Westlake Institute for Advanced Study, Hangzhou, China. [5]Westlake AI Therapeutics Lab, Westlake Laboratory of Life Sciences and Biomedicine, Hangzhou, China. [6]These authors contributed equally: Huayang Liu, Zilin Song. ✉e-mail: huangjing@westlake.edu.cn; wanghuaimin@westlake.edu.cn

https://doi.org/10.1038/s41563-025-02164-3

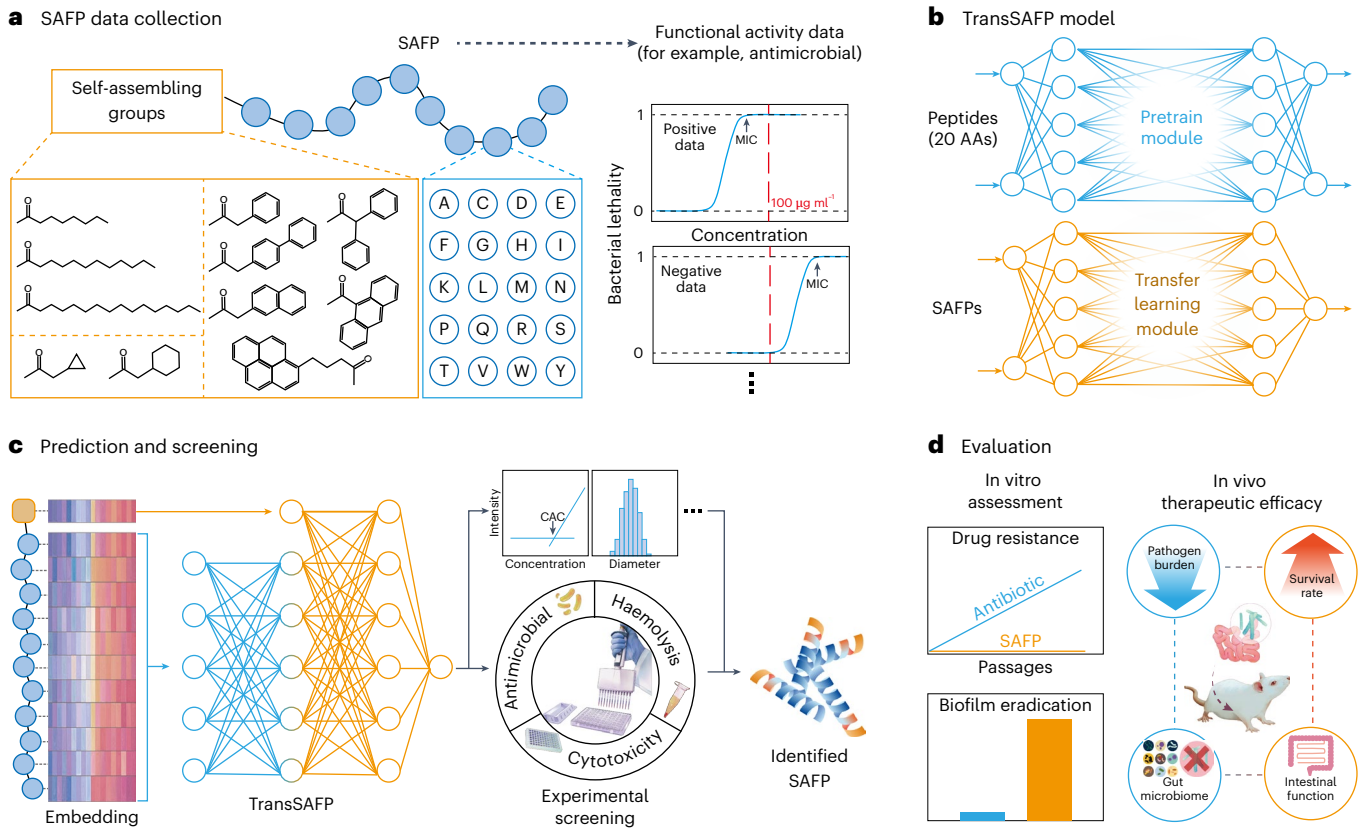


Fig. 1 | Overview of the discovery workflow of the new SAFP materials. **a**, Different types of self-assembling motifs are introduced into peptide sequences based on 20 AAs. According to the target function (for example, antimicrobial), the antimicrobial activities of SAFPs are determined via MIC experiments. The ellipsis indicates SAFPs with MIC values equal to or lower than 100 μg/ml are classified as positive data, while those with MIC values higher than 100 μg/ml are classified as negative data. The training database is compiled from publicly available data and lab data. **b**, The DL model (TransSAFP) consists of a pretrain module (blue) and a transfer learning module (orange). The pretrain module and transfer learning module are separately trained using public peptides and SAFPs, respectively. **c**, TranSAFP provides a library of SAFP candidates with antimicrobial activity. Additional experiments are carried out to characterize their self-assembling properties and screen the identified SAFPs for outstanding antimicrobial ability and biocompatibility. **d**, The identified SAFP undergoes in vitro antimicrobial assays, including the development of drug resistance and biofilm eradication. The in vivo therapeutic efficacy of the identified SAFP is evaluated by assays of survival rate, bacterial burden, gut microbiome and so on.

heavily on the sequence coevolution manifolds, which are highly noisy on the sequences with short lengths (<16 amino acids, AAs) or non-natural AAs[12], rendering the strategy unfavourable for SAFP discovery based on model fine-tuning. Progress has been made in developing specific computational tools based on the value of aggregation propensity for locating short self-assembling peptides in the vast sequence space[16–18]. However, the functional activity of these self-assembling peptide materials cannot be predicted in a straightforward manner based on computational approaches, and the strategy for effective SAFP design requires further experimental validation.

In this Article, we describe a robust and transferrable deep learning (DL) model, named the TransSAFP, which enables effective prediction of the functional activity of SAFPs relying only on minimal experimental efforts for sample annotation. We showcase the creation of new SAFPs featuring antimicrobial activity against clinically relevant bacterial strains. First, self-assembling moieties were introduced into the peptide sequences to activate self-assembling activity as a promising source of new SAFPs. The SAFPs were then subjected to minimum inhibitory concentration (MIC) tests to determine their potential antimicrobial activity with enhanced self-assembling ability (Fig. 1a). Subsequently, the TransSAFP representation learning module was initially pretrained on the public dataset of peptides composed of the 20 natural AAs. This pretraining was followed by fine-tuning the module on a downstream SAFP prediction task, incorporating self-assembling moieties and new representation augmentations (Fig. 1b). TransSAFP enables efficient high-throughput screening of the SAFPs with potent antimicrobial activity in the designated sequence space, which was corroborated by subsequent self-assembling characterizations and antimicrobial assays against multiple bacterial strains. The most potent SAFPs with antimicrobial function are further screened by biocompatibility assays, including cytotoxicity and haemolysis (Fig. 1c). The identified SAFPs were found to be unsusceptible to antimicrobial resistance and are innately capable of biofilm eradication, outperforming antibiotic agents. Finally, the identified SAFPs exhibit therapeutic efficacy in a mouse model of intestinal infection (Fig. 1d). Overall, we demonstrate the de novo design of SAFPs featuring designated bio-functions, a process that can facilitate the development of peptide materials in different scenarios.

Design of the TransSAFP model

We compiled the antimicrobial activity datasets of peptides based on 20 natural AAs from the Database of Antimicrobial Activity and Structure of Peptides (DBAASP)[19,20] (Fig. 2a). Since no data on the antimicrobial activity of SAFPs is available, we established the SAFP antimicrobial activity dataset by synthesizing peptides with self-assembling groups at the N terminus and testing their antimicrobial activities. In total, we created the SAFP antimicrobial activity dataset with balanced numbers of positive and negative training examples (Fig. 2b), which covers all 11 N-terminal types (Fig. 2c). The currently available data are yet too sparse compared to the entire (>20; ref. 15) peptide sequence space of interest,

https://doi.org/10.1038/s41563-025-02164-3

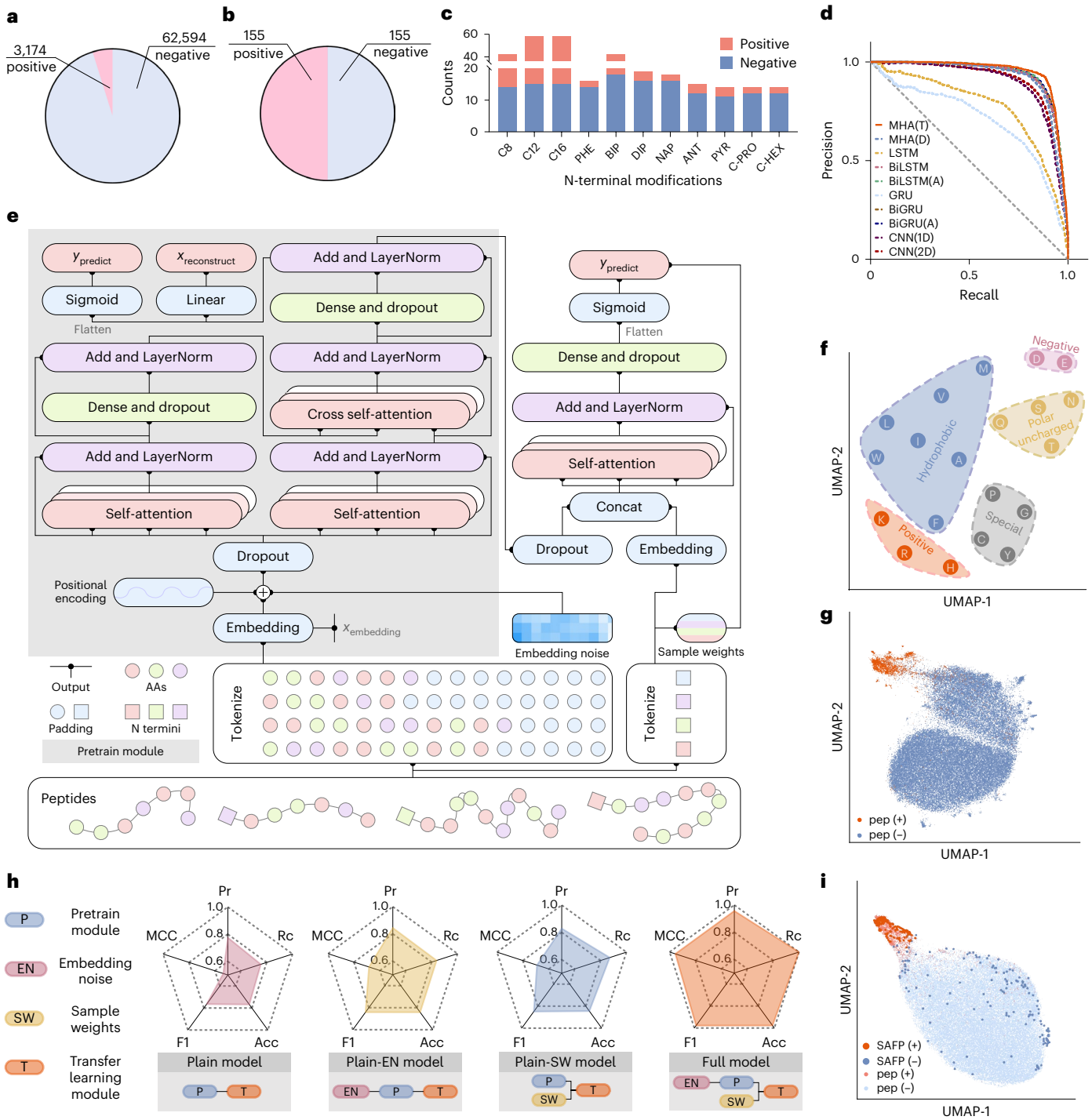


Fig. 2 | The TransSAFP protocol for identifying potential SAFPs. a,**b**, The fraction of positive and negative antimicrobial sequences in the public dataset (**a**) and the SAFP dataset (**b**). **c**, The number of antimicrobial positive and negative SAFPs classified by N-terminal modification type. **d**, The precision–recall curves of candidate sequence-learning-model architectures for the pretrain module selection. The model names are defined in the Methods. **e**, The model architecture of the pretrain–transfer learning self-attention neural network for SAFP discovery. The pretrain module (left, grey background) learns to predict the antimicrobial activity of peptides from AA sequence information and to reconstruct their sequence identity. The transfer learning module (right, white background) learns to predict the antimicrobial activity of SAFPs using the augmented latent embeddings from the pretrain module. **f**,**g**, The UMAP embeddings of the learned vectorized embedding of each AA type (**f**) and the learned latent space of the pretrain module (**g**); the labels pep (±) refer to the antimicrobially active or inactive peptides. The pretraining model exhibits high-frequency characteristics as the peptide latent features artificially aggregate into scattered clusters. **h**, The precision (Pr), recall (Rc), accuracy (Acc), F1 score (F1) and Matthews correlation coefficient (MCC) values of the transfer learning module under different augmentation schemes. From left to right are the following: the pretrain–transfer learning neural network with no augmentation (Plain model), with noise augmentation in the AA embedding space (Plain-EN model), with sample weights balancing (Plain-SW model) and with both augmentations (Full model). Note that the noise-augmented models demonstrate stronger learning power with higher scores in all metrics. **i**, The UMAP embedding of the learned latent space of the transfer learning module; the labels SAFP (±) refer to the antimicrobially active or inactive SAFPs. The high-frequency characteristics are suppressed by the sequence-invariant noise augmentation in the transfer model as the learned SAFP representations assembles two major clusters distinguishable by their antimicrobial activities with a clear decision boundary.

which we have further extended with N-terminal self-assembling groups. To overcome this challenge, we adopted the pretrain–transfer learning model, TransSAFP, which maps the input domain of natural AA sequences to the output domain of the extended SAFP space. Since the peptide sequence together with the N-terminal modification uniquely determines the properties of the SAFPs, the peptide sequences and a token for the N-terminal group are used as the input features.

We partitioned the learning task into two phases: the pretrain module and the transfer learning module. The pretrain module learns a latent representation that could correctly predict the antimicrobial labels ($y_{predict}$) and faithfully reconstructs the input sequence for downstream tasks ($x_{reconstruct}$). To begin with, we selected a multi-head attention model with a transformer-like cross-attention[21] from a benchmark test of ten classical sequence-learning-model architectures (Fig. 2d,e). The selected self-attention learning module achieved the highest performance in most metrics on the testing set and has a much faster inference time compared to other network architectures (Supplementary Figs. 1–3 and Supplementary Tables 1 and 2). In addition, directly extending the pretraining network by incorporating the self-assembling moiety tokens showed insufficient learning power on the SAFP predictions (Supplementary Fig. 4 and Supplementary Table 3**)**. This is potentially due to the fact that the natural AA peptide and the N-terminal-modified SAFPs form different distribution manifolds in the sequence space, creating the necessity of a rational training strategy for the representation learning of the SAFP feature embeddings.

We investigated the Uniform Manifold Approximation Projection (UMAP)[22] reduced token vectors of each AA from the embedding layer of the pretrain module. This analysis showed that the learned feature representations resembled the empirical AA clustering using physicochemical properties to a large extent (Fig. 2f). While the pretrain module remained accurate in predicting the antimicrobial activities, the UMAP embedding of its output latent space demonstrated rugged data distributions with negative predictions artificially aggregated into two clusters (Fig. 2g). Such typical high-frequency characteristics suggest that the pretrain module is prone to overfit on limited training samples, which consequently impedes high-quality transfer learning[23]. Furthermore, stable training of the transfer learning module is hard to realize due to the SAFPs being poisoned by unbalanced N-terminal modification distributions, especially when routine data augmentation practices are generally inapplicable to peptide encodings. However, vanilla transfer fine-tuning protocols are known to suffer from catastrophic forgetting where the established correlation between the peptide sequences and the functionalities is discarded by the downstream learning module, such that the learned representation in the pretraining displays no increment for the transfer learning task.

Considering these factors, we designed the transfer learning module to extend the pretrain sequence embeddings towards the SAFP input domain (Fig. 2e). Specifically, the transfer learning module inputs were constructed by concatenating the latent output from the frozen pretrain module with the N-terminal self-assembling groups (or no self-assembling group) using a separated embedding entry. The concatenated latent features were embedded through a standard self-attention block and were flattened to predict the final antimicrobial activity label via a sigmoid activation. To preserve the learned correlation in the pretrain module, we also incorporated the peptide sequences from the pretraining natural AA sequences to form the SAFP dataset. Since the SAFP dataset was swarmed with the public dataset, the transfer learning module was biased towards the SAFPs by assigning sample weights, w_j, to the loss function according to the occurrence of the jth type of N-terminal self-assembling group in the dataset. In this case, the loss contributions of the public non-modified sequences act essentially as regularization terms aiming to prevent the catastrophic forgetting of AA sequence contributions in the downstream module (Supplementary Fig. 5 and Supplementary Table 3). Notably, a noise augmentation scheme was introduced for the AA token vectors in the peptide sequence embedding input. Specifically, during the transfer learning module training, we introduced noise vectors sampled uniformly from the spherical tessellations of the pretrain embedding space. The noise augmentations introduced here are sequence immutable as guaranteed by the noise sampling bounds and computationally efficient due to the number of AA embeddings (Methods). The transfer learning module trained with both sample weight biasing and noise-augmented embedding could correctly recognize all antimicrobial SAFPs with only 7 false positives out of 310 predictions. While the full model (TransSAFP) outperforms the augmentation-ablated models (Fig. 2h) as well as models from alternative training strategies (Supplementary Figs. 4 and 5 and Supplementary Table 3) in all performance metrics, we note that our noise augmentations to the embedding annotation largely improve the models' performance in both the public dataset and the SAFP dataset (Supplementary Figs. 6 and 7 and Supplementary Table 4). In principle, the noise embedding smooths the discrete feature space formed by the quantized token vectors and suppresses the high-frequency characters on the loss landscape, which allows convergence to a flat minimum and promotes model generalization. Moreover, the UMAP embedding of the latent space in the full transfer learning module demonstrates an ordered learned distribution with a smooth decision boundary that differentiates the antimicrobial active and non-active SAFPs (Fig. 2i). Given the minimal scope for further improvement in the present transfer module, we proceeded with the current TransSAFP protocol for the subsequent screening of potential antimicrobial SAFPs.

TransSAFP-guided screening of antimicrobial SAFPs

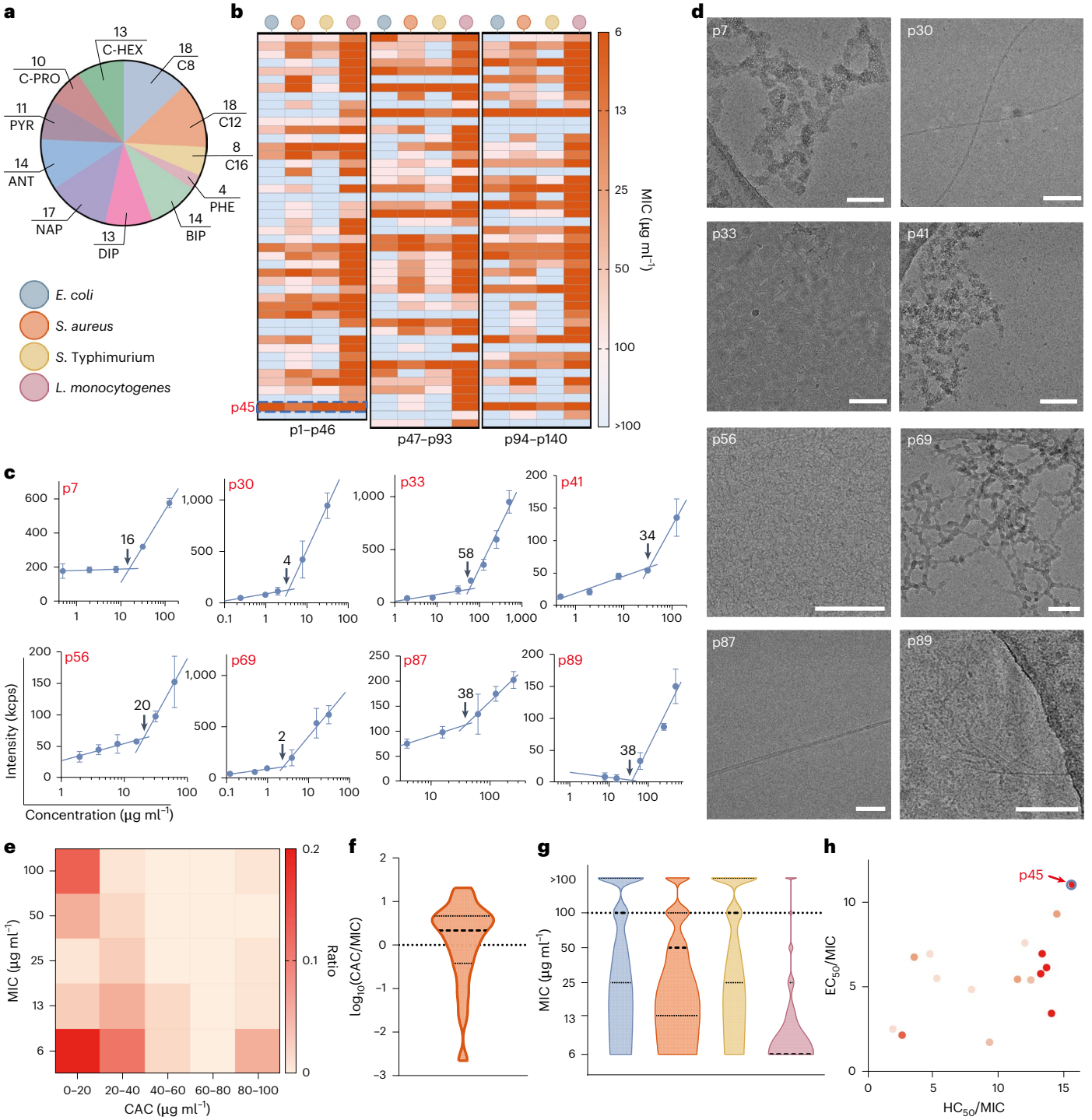
First we used peptide sequences from the public dataset and 11 self-assembling groups to generate a new SAFP sequence space (length of sequences, 6–15 AAs) in silico and then employed TransSAFP to predict it. Subsequently, we shortlisted the SAFP sequences that achieved high model prediction scores (>0.99) for their antimicrobial activity. From this shortlist, we selected 140 SAFP candidates (labeled as p1–p140, Supplementary Table 5) with a balanced proportion of all N-terminal types (Fig. 3a), which are subjected to subsequent

Fig. 3 | Experimental validation and screening of SAFP candidates. **a**, Decomposition of 140 chemically synthesized SAFP candidates by the number (above line) of each N-terminal type (below line). **b**, MIC assays against *E. coli*, *S. aureus*, *S.* Typhimurium and *L. monocytogenes* confirmed that 121 out of 140 SAFP candidates exhibited antimicrobial activity, demonstrating a success rate of ~86% for designing antimicrobial SAFPs with TransSAFP. **c**, Identified SAFPs exhibited a self-assembling ability, confirmed by CAC measurements. CAC values are indicated by black arrows. The bars shown are mean ± s.d. ($n = 3$ independent replicates). **d**, Cryo-EM images of antimicrobial SAFPs demonstrate that the identified SAFPs self-assemble into structures with distinct morphologies. Scale bars, 100 nm. **e**, The ratio of synthesized SAFP candidates with different antimicrobial activities in different CAC ranges, showing that stronger self-assembly ability is generally associated with enhanced antimicrobial activity. **f**, The $\log_{10}$(CAC/MIC) value distribution of SAFP candidates reveals that most SAFPs exert antimicrobial activity at concentrations above or near their CAC values. The dotted lines represent the quartiles and the dashed lines represent the median, while the line at zero is a guide for the eye. **g**, MIC distributions of all chemically synthesized SAFP candidates. Identified SAFPs show good activity against four bacterial strains. The figure legend is same as that for **b**. The dotted lines represent the quartiles and the dashed lines represent the median, while the line at 100 is a guide for the eye. **h**, EC_{50}/MIC and HC_{50}/MIC selective indexes on highly potent antimicrobial SAFP candidates. The p45 exhibits the best selectivity among identified SAFPs. HC_{50} and EC_{50} were determined from rabbit red blood cells and GES-1 cells, respectively. The red colour map illustrates the MIC values against *S.* Typhimurium, the same as in **b**. **i**, MICs of p45 against 14 bacterial strains, showing its broad-spectrum antimicrobial activity.

https://doi.org/10.1038/s41563-025-02164-3

chemical synthesis and purification. MIC assays against four bacteria strains (*Escherichia coli*, American Type Culture Collection (ATCC) 25922; *Staphylococcus aureus*, ATCC 25923; *Salmonella enterica* subsp. *enterica serovar* Typhimurium, SL1344; and *Listeria monocytogenes*, National Center for Medical Culture Collections (CMCC) 54004) were conducted to validate the antimicrobial activity of the synthesized candidates (Fig. 3b).

We identified 121 SAFP candidates that are inhibitive to at least one of the pathogen strains with MICs of ≤100 μg ml^{-1}, corresponding to an ~86% success rate of antimicrobial SAFP design in the low-data SAFP regime. Simultaneously, we note that the N-terminal self-assembling moieties can enhance the antimicrobial activity of some peptides or transform non-antimicrobial peptides (non-AMPs) into antimicrobial ones. To elucidate the underlying mechanisms of the self-assembly, we first simulated the selected SAFPs using coarse-grained molecular dynamics in explicit solvent. The coarse-grained molecular dynamics results demonstrate that the SAFPs self-assemble within the 500 ns simulation time with an aggregation propensity of >1, suggesting



i

	ATCC 25922	ATCC 25923	SL 1344	CMCC 54004	USA 300	BNCC 186732	BNCC 363037	ATCC 51559	ATCC 51299	ATCC 51575	BNCC 330867	BNCC 254392	BNCC 102997	BNCC 360085
p45 (μg ml^{-1})	6	13	6	6	13	13	13	13	100	100	6	6	6	13

theoretically their strong tendency to auto-aggregate (Supplementary Figs. 8 and 9). The self-assembling ability of the SAFPs was evaluated by assays in critical aggregation concentrations (CACs). The CAC values of most antimicrobial SAFPs are less than 100 μg ml^{-1} (Fig. 3c and Supplementary Fig. 10), validating their strong self-assembling ability in solution. Dynamic light scattering (DLS) measurements indicate that the discovered SAFPs can form nanostructures in solution, as evidenced by the correlation function (Supplementary Fig. 11). Cryogenic electron microscopy (cryo-EM) images (Fig. 3d and Supplementary Fig. 10) reveal that the SAFPs exhibit a diverse range of assembly morphologies, including nano aggregates (p7, p38, p41, p45), long fibres (p27, p30, p87), short fibres (p89, p109), worm-like structures (p33, p56), nanonets (p69, p97, p103, p124) and sheets (p73). The reported SAFPs in Fig. 3d displayed at least two types of secondary structure after self-assembly, typically dominated by random coil structure with a small fraction of β-sheet or α-helix, as observed with circular dichroism (CD) and Fourier transform infrared (FTIR) spectroscopies (Supplementary Fig. 12 and Supplementary Tables 6 and 7). Due to the different sample conditions in CD and FTIR measurements (Supplementary Information), slight discrepancies exist in the analysis results. However, the CD analysis should provide the in situ secondary structure of the reported SAFPs in phosphate-buffered saline (PBS) buffer. We observed downshifts on the C=O group and the broader peaks of the NH group, which indicate the formation of stronger hydrogen bonds in SAFP assemblies[24–26] (Supplementary Figs. 13 and 14 and Supplementary Tables 8 and 9). Perturbation reflected by the shift of ethylene C–H stretches indicates the existence of interactions to alkyl-chain N-terminal modifications in p7, p30, p87 and p89 (ref. 27). Synergistically, wide-angle X-ray scattering measurements (Supplementary Fig. 15) displayed the characteristic peak for $\pi-\pi$ staking at the d-spacing $d_{\pi-\pi}$ = 3.4 Å (ref. 28) in SAFP assemblies with aromatic N terminals (p33, p41, p56 and p69). This phenomenon was also observed in p7, which is potentially attributed to the presence of the aromatic AA (tryptophan) within its sequence. Overall, the integrated analysis based on DLS, cryo-EM, CD spectroscopy, FTIR spectroscopy and wide-angle X-ray scattering provides a comprehensive understanding of the structural features and intermolecular interactions driving the self-assembly of the discovered SAFPs.

As shown in Fig. 3e, the antimicrobial activity of these SAFPs is related to their self-assembling ability as more than half of SAFPs are distributed in the lower-left region, indicating that the SAFPs with a stronger ability to self-assemble often exhibit more excellent antimicrobial activity. Approximately 40% of the CAC values are lower than the MIC values of SAFPs ($\log_{10}$(CAC/MIC) < 0), suggesting the formation of self-assemblies during the antimicrobial function of the SAFP (Fig. 3f). In principle, peptides generally can accumulate near bacteria, leading to a localized increase in concentration, reaching the CAC in specific regions where SAFPs form self-assemblies and exert antimicrobial function, which explains the MIC values being slightly lower than the corresponding CAC measures for the rest of the SAFPs. To further investigate the potential connection between the self-assembling property of the discovered SAFPs and their antimicrobial activity, we synthesized multiple SAFP analogues with similar physicochemical properties but distinct self-assembling properties. By retaining the AA composition and other key characteristics, we aimed to investigate the effect of self-assembly on the antimicrobial function of the discovered SAFPs. The results (Supplementary Table 10) demonstrate that when the physicochemical properties are consistent, the peptides with a stronger self-assembling propensity, characterized by lower CACs, generally display enhanced antimicrobial activity with lower MICs, and vice versa. Furthermore, we found that disrupting the self-assembled structures of the SAFPs using the surfactant Tween-80 leads to a loss of their antimicrobial activity at concentrations where the peptides would otherwise self-assemble and exhibit antimicrobial effects (Supplementary Fig. 16). In summary, these results demonstrate that the self-assembling ability of the discovered SAFPs typically determines their antimicrobial activity.

The MIC value distribution (Fig. 3g) indicated that Gram-negative bacteria, and *S.* Typhimurium in particular, are less susceptible to most designed antimicrobial SAFPs. Therefore, we referred to this strain as the potency measure for subsequent antimicrobial SAFP screening based on two selective indexes, HC_{50}/MIC and EC_{50}/MIC, which were derived from cytotoxicity, haemolysis and MIC experiments (HC_{50} is the concentration needed to cause 50% haemolysis; EC_{50} is the 'effective concentration' or the concentration of the antimicrobial that inhibits 50% cellular growth). The SAFP p45 was selected as the most prominent antimicrobial candidate due to its outstanding biocompatibility (Fig. 3h and Supplementary Table 11). Moreover, p45 also showed broad-spectrum antimicrobial capability against ten additional pathogenic strains (Fig. 3i), including *Enterobacter cancerogenus* (BeNa Culture Collection (BNCC) 363037), *Staphylococcus epidemidis* (BNCC 330867), *Acinetobacter baumannii* (BNCC 254392), *Klebsiella pneumoniae* (BNCC 102997) and *Pseudomonas aeruginosa* (BNCC 360085). Moreover, p45 also exhibited good efficacy against drug-resistant strains, such as *Staphylococcus aureus* (USA 300, methicillin resistant), *E. coli* (BNCC 186732, multidrug resistant), *Enterococcus faecium* (ATCC 51559, multidrug resistant), *Enterococcus faecalis* (ATCC 51299, vancomycin resistant) and *Enterococcus faecalis* (ATCC 51575, multidrug resistant). In particular, p45 exhibits good antimicrobial activity against several ESKAPE pathogens.

To identify antimicrobial SAFPs with AA sequences that are far from the known peptide space, we also adopted a de novo screening protocol that focused on the entire octa-SAFP library with all 11 types of N-terminal modifications (Fig. 4a; Methods). Most of the filtered antimicrobial octa-SAFP candidates have at most two AAs aligned to one of the known AMP sequences, with the similarity level mostly below 30% (Fig. 4b). Figure 4c illustrates the difference of the TransSAFP prediction scores with and without the N-terminal moiety inputs for the selected octapeptide sequences. The positive score increments from octanoic acid (abbreviated as C8), lauric acid (C12), hexadecanoic acid (C16), 4-biphenylacetic acid (BIP) and 1-pyrenebutyric acid (PYR) modifications in this sequence space indicate that these N-terminal modifications are more likely to enhance the antimicrobial activity of the dissimilarity-filtered octapeptide sequences, which aligns with higher portions occupied by these N-terminal modifications in the region of high TransSAFP scores (>0.99; Fig. 4d).

We further compared the properties of predicted octa-SAFP candidates with known AMP sequences, including charge, hydrophobicity, hydrophobic ratio, aromaticity and AA composition (Fig. 4e–i). The analysis was carried out based on the different N-terminal groups, including alkanes (C8, C12 and C16), aromatic rings (phenylacetic acid, PHE; BIP; diphenylacetic acid, DIP; 2-naphthaleneacetic acid, NAP; 9-anthracenecarboxylic acid, ANT; and PYR) and cycloalkanes (cyclopropylacetic acid, C-PRO and cyclohexaneacetic acid, C-HEX). The discovered octa-SAFPs have similar charges to the known AMPs from the training dataset, but the charges of octa-SAFPs with alkane and aromatic N-terminal modifications are slightly lower than those in other groups (Fig. 4e). The hydrophobicity of 20 common AAs and 11 N-terminal modifications was re-quantified by the Kovacs scale[29,30] (Supplementary Table 12) to calculate the hydrophobicity of peptides, as shown in Fig. 4f. A notable increase in the hydrophobicity of SAFPs is observed and is related to the presence of the N-terminal modifications. Alkane showed the greatest increase, followed by aromatic rings, and cycloalkane showed the least increase. Due to the considerable hydrophobicity provided by the N-terminal modifications, the hydrophobic ratio of SAFPs has notably decreased compared to known AMPs (Fig. 4g). Interestingly, TransSAFP increases the aromaticity in predicted octa-SAFPs (Fig. 4h). Regarding the changes in the AA compositions, we found that the introduction of hydrophobic N-terminal modifications can substantially reduce the hydrophobicity of the AA

https://doi.org/10.1038/s41563-025-02164-3

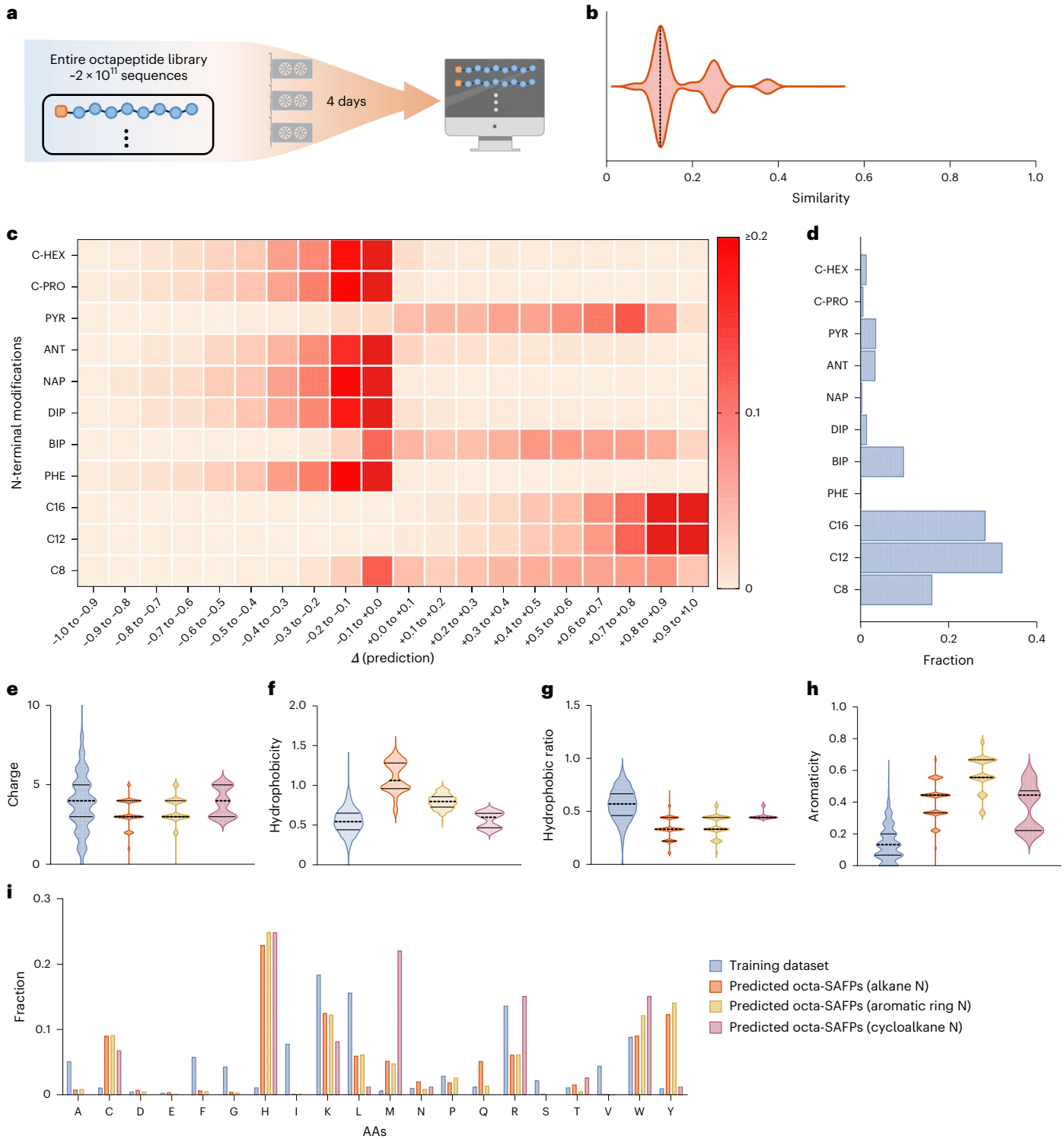


Fig. 4 | Prediction and analysis of SAFP in entire octapeptide library. **a**, TransSAFP can screen the entire octapeptide library in four days. **b**, Predicted octa-SAFP candidates exhibit low similarity to known AMP sequences from the training dataset. The dashed line represents the median. **c**, Distribution of the changes in prediction scores (*Δ*) with various N-terminal moieties. The colour bar represents the ratio values. Among these N-terminal moieties, C8, C12, C16, BIP and PYR exhibit a higher likelihood of enhancing the antimicrobial activity of the octapeptides. **d**, Distribution of different N-terminal moieties in the octa-SAFP dataset, with prediction values ranging from 0.99 to 1.00. C8, C12 and C16 constitute the top proportions, followed by BIP, PYR and ANT. **e**–**h**, Physicochemical properties of octa-SAFP candidates and known AMP sequences from the training dataset, including charge (**e**), hydrophobicity (**f**), hydrophobic ratio (**g**) and aromaticity (**h**). The dotted lines represent the quartiles and the dashed lines represent the median. **i**, AA composition of predicted octa-SAFP and known AMP sequences from the training dataset. Panels **e**–**i** reveal that the predicted octa-SAFPs exhibit notable differences in physicochemical properties and AA composition compared to known AMP sequences. The label 'alkane N' includes the N-terminal modifications C8, C12 and C16; 'aromatic ring N' includes PHE, BIP, DIP, NAP, ANT and PYR; and 'cycloalkane N' includes C-PRO and C-HEX.

sequences (Supplementary Fig. 17). Besides, the ratios of K and R in the predicted octa-SAFP candidates decrease, while a substantial increase in H occurs, resulting in a similar charge distribution compared to known AMP sequences (Fig. 4i). Due to the presence of hydrophobic N-terminal modifications in SAFPs, the ratio of hydrophobic AAs (such as F, G, I and L) decreases notably, leading to a lower hydrophobic ratio of SAFP sequences. The increase in the ratio of H and Y causes an elevation in aromaticity. Compared to octa-SAFPs with N-terminal modifications of alkanes or aromatic rings, octa-SAFPs with cycloalkane require more hydrophobic AAs such as M and W, as well as the positively charged R, to facilitate antimicrobial activity. We synthesized a batch of octa-SAFPs (sequences are listed in Supplementary Table 5), tested their antimicrobial activity against the four aforenoted pathogen strains and found that octa-p2 exhibited the best antimicrobial activity (Supplementary Table 13). The successful identification of multiple antimicrobial octa-SAFPs with almost no similarity to previously known AMP sequences validates the TransSAFP-guided discovery of new SAFPs with non-natural AAs. This promises compelling opportunities to develop next-generation antimicrobial therapeutic agents in an unexplored sequence space.

In vivo toxicity assay

Prior to evaluating the therapeutic efficiency of identified SAFPs, we first systematically examined their biosafety (Supplementary Fig. 18a). In blood biochemical analysis (Supplementary Fig. 18b), the uric acid (UA) of p45-treated mice was significantly higher than in the PBS group, while other parameters (alanine transaminase, aspartate aminotransferase, blood urea nitrogen and creatinine) remained relatively stable upon intraperitoneal injection treatments. However, haematological evaluation (Supplementary Fig. 18c) exhibited no significant difference in haematological indicators, including red blood cell, mean corpuscular volume, platelet, white blood cell, neutrophil and monocyte measures. Haematoxylin and eosin (H&E) staining results suggested that no notable abnormalities are observed from the heart, liver, spleen, lung, kidney, small intestine or colon (Supplementary Fig. 19). In the subsequent five days, the body weight (Supplementary Fig. 20a) of the mice decreased slightly after the treatment of ciprofloxacin or two SAFPs on day 2. Mice treated with ciprofloxacin and octa-p2 began to recover within a day, while p45-treated mice showed a slower recovery. However, on day 7, the body weight of all groups returned to normal levels. The chronic toxicity from assessment experiments (day 7) suggested that no statistical difference exists in blood biochemical or haematological indicators among treated mice groups (Supplementary Fig. 20b,c). Simultaneously, no apparent abnormalities were shown in tissue slices (Supplementary Fig. 21). These results reveal that the two identified SAFPs could result in a slight decrease in body weight, and p45 would cause a burden to the kidneys of mice. Nevertheless, these effects are temporally transient, and relevant indicators will revert within five days, indicating no long-term damage was caused by the two identified SAFPs.

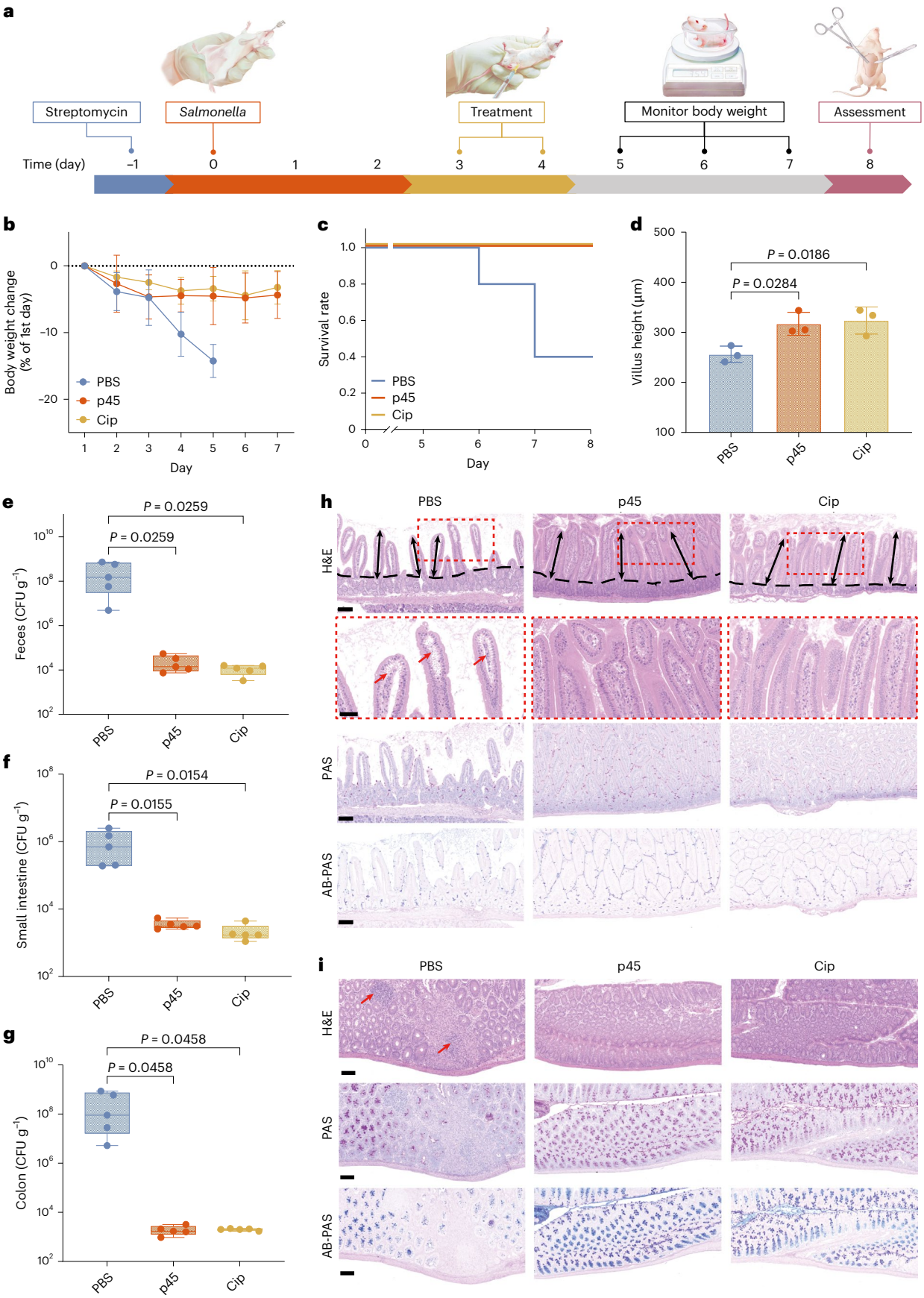
In vivo therapeutic efficacy on intestinal infection

The therapeutic efficacies of SAFPs via intraperitoneal injection were subjected to in vivo assessment in an acute intestinal infection mouse model developed by *S.* Typhimurium (Fig. 5a). The body weight of the infected mice treated with PBS consistently decreased (Fig. 5b), and some reached the threshold of losing 20% (humane end-point) of initial body weight after day 5. Treatment with p45 could prevent weight loss, similar to ciprofloxacin, and both of them ensured a high survival rate of infected mice compared to the PBS group (Fig. 5b,c). The p45 therapy leads to a decrease of *Salmonella* burden in faeces, the small intestine and the colon, with efficacy comparable to that of ciprofloxacin (Fig. 5e–g). The *S.* Typhimurium infection also inflicted severe damage upon the intestinal tissues, including a significant decrease in the density and height and the swelling of small intestinal villi (Fig. 5d,h). The numbers of goblet cells on the villi of the small intestine in the p45-treated and ciprofloxacin-treated groups are notably higher than that in the PBS group, as illustrated in periodic acid Schiff (PAS) and Alcian blue PAS (AB-PAS) stained images (Fig. 5h). The thickening of the muscular layer, structural disorders and inflammatory region in the colon were observed in the PBS group (Fig. 5i). By contrast, these pathological symptoms were greatly alleviated upon treatment of p45 or ciprofloxacin.

To investigate the changes in the relative abundance of the gut microbiome induced by p45 and ciprofloxacin treatments, we further profiled the average bacterial taxonomy using 16S ribosomal RNA (rRNA) analysis. The unweighted pair group method with arithmetic mean tree plot (Supplementary Fig. 22a) showed that most p45 and ciprofloxacin samples were grouped together, and most PBS samples were clustered into another clade. Furthermore, the diversity of the bacterial community between different treated mice was also analysed by weighted UniFrac principal coordinates analysis (Supplementary Fig. 22b) and non-metric multidimensional scaling (Supplementary Fig. 22c). Both of these also showed that mice treated with p45 distributed densely and can be distinguished from the PBS group, while the ciprofloxacin-treated group exhibits widely scattered points in both analysis plots, indicating that p45 treatment could lead to more consistent gut microbiota, as compared to ciprofloxacin. Average bacterial taxonomic profiling plots provide the top ten relative abundance at different levels. The p45 and ciprofloxacin treatments greatly increase the Firmicutes and Bacteroidota phyla compared to the PBS group (Supplementary Fig. 22d). The abundance of the Clostridia and Bacteroidia classes increased notably after treatment with p45 or ciprofloxacin compared to the PBS-treated mice (Supplementary Fig. 22e). The Clostridia class is involved in processes that maintain gut homeostasis. For example, the *Clostridium* species produce short-chain fatty acids, which are beneficial for colonic health[31]. Members of the Bacteroidia class play vital roles in the metabolization of oligosaccharides and polysaccharides, thus supplying nutrition not only to the host but also to other commensal microbiota[32]. Notably, ciprofloxacin treatments increased the abundance of the Bacilli class, which contains some

Fig. 5 | Therapeutic efficacy of p45 against intestinal infection. a, The protocol of in vivo therapeutic assay, including streptomycin pretreatment, *Salmonella* infection, treatments, body weight monitoring and final assessment. **b**, Body weight change of mice groups treated with PBS, p45 and ciprofloxacin (Cip). Compared to the control group (PBS treatment), p45 and Cip treatment effectively prevented weight loss. The bars shown are mean ± s.d. ($n = 5$ biologically independent samples). **c**, Both p45 and Cip effectively ensured the survival rate of mice with acute intestinal infection caused by *S.* Typhimurium ($n = 5$ biologically independent samples). **d**, Average villus height measured from H&E-stained images of small intestines from different groups. The bars shown are mean ± s.d. The difference between PBS and other treatments is determined by the one-way analysis of variance (ANOVA) test. **e**–**g**, *S.* Typhimurium quantity in the faeces (**e**), small intestine (**f**) and colon (**g**) of infected mice with different treatments. In **e**–**g**, five horizontal solid lines in each dataset represent the following, from top to bottom: maximum (the top short line), first quartile (the top of the box), median (the line within the box), third quartile (the bottom of the box) and minimum (the bottom short line). The difference between PBS and other treatments is determined by the one-way ANOVA test ($n = 5$ biologically independent samples). **h**, Sections of small intestine tissue from infected mice with different treatments. The first row shows H&E staining; scale bar, 100 μm. The second row shows a zoomed-in view of the red dashed boxes; scale bar, 50 μm. The third row shows PAS staining; scale bar, 100 μm. The fourth row shows AB-PAS staining; scale bar, 100 μm. The heights of villi were measured as the length of the black arrows. The swollen areas of the villi are marked with red arrows. **i**, Sections of colon tissue from infected mice with different treatments. The first row shows H&E staining; scale bar, 100 μm. The second row shows PAS staining; scale bar, 100 μm. The third row shows AB-PAS staining; scale bar, 100 μm. The inflammatory regions are marked with red arrows. Panels **h** and **i** reveal that mice treated with p45 and Cip showed better small intestine and colon conditions compared to the PBS group.

https://doi.org/10.1038/s41563-025-02164-3



https://doi.org/10.1038/s41563-025-02164-3

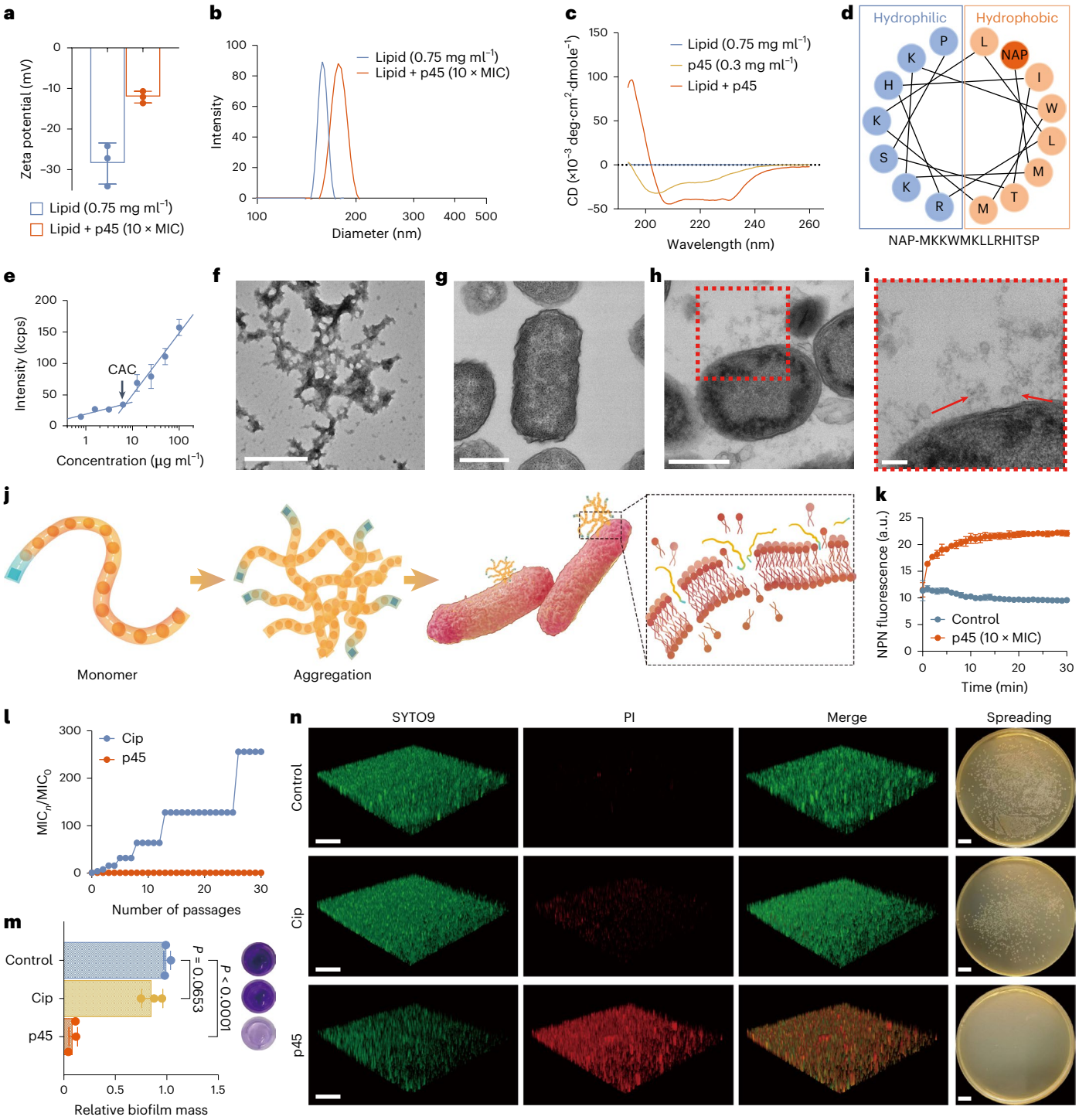


Fig. 6 | Antimicrobial mechanisms and biofilm eradication of p45. **a**–**c**, The zeta potential (**a**), diameter distribution (**b**) and CD spectrum (**c**) of lipid vesicles (DOPC/DOPG = 7:3) before and after treatment with p45, revealing the interaction between lipid vesicles and p45. **d**, The sequence and helical wheel of p45. **e**, The p45 exhibits a strong self-assembly ability with a CAC value of 6 μg ml^{-1}. kcps, kilo counts per second. **f**, TEM examination of p45 (10 × MIC); scale bar, 500 nm. **g**,**h**, TEM images of *S.* Typhimurium cells (**g**) and cells treated with p45 (50 × MIC) for 15 min (**h**); scale bars, 500 nm. **i**, Zoomed-in view marked by the red dashed box in **h**; scale bar, 100 nm. Self-assemblies are marked by red arrows. Panels **g**–**i** reveal that p45 assemblies target and bind to the bacterial membrane surface. **j**, Illustration of antimicrobial mechanism of p45. **k**, The p45 disrupts the integrity of the outer membrane of *S.* Typhimurium, leading to an increase in NPN signal. a.u., arbitrary units. **l**, Compared to Cip, p45 does not induce *S.* Typhimurium resistance within 30 passages. MIC_0 is the initial MIC value, while MIC_n is that value after *n* passages, as described in more detail in the Methods. **m**, Relative mass and crystal-violet-stained images (inset) of *S.* Typhimurium biofilm treated with PBS (control), Cip (50 × MIC) and p45 (50 × MIC). The bars shown are mean ± s.d. The difference between PBS and other treatments are determined by the one-way ANOVA test. **n**, CLSM images (scale bar, 100 μm) and spreading plate results (scale bar, 1 cm) of established *S.* Typhimurium biofilm treated with PBS (control), Cip and p45. Panels **m** and **n** indicate that p45 exhibits a better biofilm eradication ability compared to Cip. In **a**, **e**, **k** and **m**, the bars shown are mean ± s.d. ($n = 3$ independent replicates).

opportunistic pathogens. However, such side effects are not found in mice treated with p45 (ref. 33). At the family level, the variance of abundance in Lachnospiraceae and Muribaculaceae is not notable (Supplementary Fig. 22f). However, the abundance of the Proteobacteria phylum (Supplementary Fig. 22g), Gammaproteobacteria class (Supplementary Fig. 22h) and Enterobacteriaceae family (Supplementary Fig. 22i) containing several *Salmonella* species is suppressed after treatment with p45 or ciprofloxacin, indicating that p45 exhibits a similar therapeutic efficacy to ciprofloxacin in eliminating in vivo bacteria.

Furthermore, experiments were also carried out to evaluate the in vivo therapeutic efficacy of octa-p2. The results show that treatment of octa-p2 could effectively prevent weight loss and eradicate *S.* Typhimurium, ensuring the survival of infected mice (Supplementary Fig. 23a–c). Upon treatment, infection symptoms were alleviated, including the abnormal status of small intestinal villi and the disordered structure of the colon (Supplementary Fig. 23d,e). From 16S rRNA analysis at the class level (Supplementary Fig. 23f), treatment with octa-p2 reduced the abundance of the Gammaproteobacteria class and promoted the population of the Clostridia class.

Antimicrobial mechanism of SAFP p45

To unravel the antimicrobial mechanism of the identified SAFP (p45), we employed experiments at both the membrane and cellular levels. A lipid bilayer model was constructed by small unilamellar vesicles (DOPC/DOPG = 7:3; DOPC, 1,2-dioleoyl-*sn*-glycero-3-phosphocholine; DOPG, 1,2-dioleoyl-*sn*-glycero-3-phospho-(1′-rac-glycerol) (sodium salt)) to investigate the interaction between p45 and lipid membranes. After interaction with p45 at 10 × MIC, the values of the zeta potential and diameter of small unilamellar vesicles increased (Fig. 6a,b), indicating the binding of p45 to the small unilamellar vesicle membranes. CD results (Fig. 6c and Supplementary Table 6) showed that p45 itself exhibited a mixed secondary structure in PBS buffer, which transformed into a helix-dominated and more amphiphilic structure (Fig. 6d) after interaction with small unilamellar vesicles.

The CAC of p45 was determined using DLS technology, and its CAC value of 6 μg ml^{-1} (3 μM) is equivalent to its MIC against *S.* Typhimurium (Fig. 6e). Therefore, p45 has already undergone self-assembly at an antimicrobial concentration. Moreover, when the p45 assemblies are disturbed at a concentration higher than its MIC, p45 loses its antimicrobial activity (Supplementary Fig. 16e), which indicates that p45 first forms assemblies before binding to the bacterial membranes and exerting antimicrobial effects. Transmission electron microscopy (TEM; Fig. 6f) and atomic force microscopy (AFM; Supplementary Fig. 24b) images indicate that p45 self-assembled to form a nanofibrous aggregation. To mitigate the potential interferences in the TEM and AFM techniques, including uranyl acetate negative staining and dry environment, we further used DLS to investigate the correlation function of p45 assemblies in solution (Supplementary Fig. 24a), which indicates the formation of a nanostructure of p45. Additionally, the cryo-EM image of p45 reveals a consistent morphology to that observed from TEM and AFM (Supplementary Fig. 24c). The cell membrane of healthy *S.* Typhimurium exhibited an intact structure (Fig. 6g). After a 15 min interaction with p45, TEM results (Fig. 6h,i) revealed the binding of p45 assemblies to the bacterial membrane, which is consistent with the results from the small unilamellar vesicle experiments. After 1 h treatment, the bacterial membrane was disrupted and the cytoplasm leaked, resulting in the death of bacteria (Supplementary Fig. 25a,b). To further investigate the antimicrobial mode of p45 assemblies in living bacterial cells, we used 3,3′-dipropylthiadicarbocyanine iodide ($DiSC_3(5)$) and *N*-phenyl-1-naphthylamine (NPN) fluorescence dyes to examine the changes in bacterial membrane depolarization and permeability. In the $DiSC_3(5)$ assay (Supplementary Fig. 26), unlike the well-known depolarizing peptide antibiotic polymyxin B[34], no increase of $DiSC_3(5)$ fluorescence was observed upon the treatment of p45, indicating that the antimicrobial mechanism of p45 is independent of depolarization to the membrane. NPN experiments showed the fluorescent signal increased substantially after contact with p45 (Fig. 6k), indicating the permeability change of the membrane. Overall, these results imply that p45 can self-assemble into an aggregation below its MIC and the aggregation subsequently binds to the bacterial membrane, leading to a decrease in permeability and disruption in the integrity of the bacterial membrane (Fig. 6j).

Development of resistance and biofilm eradication

The development of resistance experiment (Fig. 6l) indicates that *S.* Typhimurium did not develop acquired resistance against p45 after 30 passages. By contrast, the MIC of ciprofloxacin increased by 256 times compared to the initial value.

The biofilm eradication assays show that the established biofilm was almost entirely eradicated (Fig. 6m) after a 4 h treatment of p45 (50 × MIC), while approximately 90% of the biofilm remained upon treatment with ciprofloxacin. Confocal laser scanning microscopy (CLSM) images by live/dead assay of the biofilm provide a direct observation of the viability of bacteria in the biofilm upon the treatment with p45 or ciprofloxacin. The results suggest (Fig. 6n) that ciprofloxacin can eliminate only a small portion of bacteria in the biofilm. By sharp contrast, CLSM images and spreading plate results indicate that almost all bacteria within the biofilm treated with p45 have died. Since biofilm formation is an important factor in antibiotic resistance and infection recurrence, our results suggest that p45 has advantages in these aspects.

Outlook

Peptide materials with diverse bio-functionalities have potential applications in various fields. However, the lack of a computational tool for predicting the functional activities of SAFPs dramatically limits the development of new SAFPs. This work proposed a DL-assisted pipeline (TransSAFP) for designing N-terminal-modified SAFPs with antimicrobial function. We show that, with appropriate augmentation schemes in the transfer learning phase, the TransSAFP model accurately predicts the antimicrobial activities of SAFPs by learning on sparsely labelled data in the extended chemical-sequence space. TransSAFP demonstrates an effective scheme for antimicrobial SAFP discovery, as verified by MIC assays. The identified SAFPs strongly self-assemble into diverse morphologies with various secondary structures. N-terminal modifications enhance self-assembly via hydrogen bonds, hydrophobic interactions and π–π stacking, and stronger self-assembly generally leads to greater antimicrobial activity.

The identified SAFPs (p45 and octa-p2) demonstrated good biocompatibility in vitro and in vivo, showing effective therapeutic efficacy against acute intestinal bacterial infections. Importantly, the SAFPs were able to restore the intestinal environment and maintain the gut microbiota, comparable to the effects of the antibiotic ciprofloxacin. The p45 exhibits a strong self-assembling ability and adopts a helix-dominated secondary structure upon interacting with the bacterial membrane. The nanofibrous aggregation of p45 attached to the membrane and permeabilized it without leading to depolarization. With such an antimicrobial mechanism, p45 exhibited an excellent biofilm eradication ability and broad-spectrum antimicrobial activity without inducing drug resistance, demonstrating greater potential in clinical treatments for bacterial infections. To conclude, we developed a transfer learning DL strategy (TransSAFP) for the de novo design of functional peptide-based materials. The TransSAFP-assisted workflow presented here can effectively be adapted to design other SAFP materials with various bio-functionalities.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions

and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41563-025-02164-3.

References

1. Ahnert, S. E. et al. Principles of assembly reveal a periodic table of protein complexes. *Science* **350**, aa2245 (2015).
2. Vermeire, P.-J. et al. Molecular interactions driving intermediate filament assembly. *Cells* **10**, 2457 (2021).
3. Yang, G. et al. Precise and reversible protein-microtubule-like structure with helicity driven by dual supramolecular interactions. *J. Am. Chem. Soc.* **138**, 1932–1937 (2016).
4. Imada, K. Bacterial flagellar axial structure and its construction. *Biophys. Rev.* **10**, 559–570 (2018).
5. Jia, Y. & Li, J. Molecular assembly of rotary and linear motor proteins. *Accounts Chem. Res.* **52**, 1623–1631 (2019).
6. Chiesa, G., Kiriakov, S. & Khalil, A. S. Protein assembly systems in natural and synthetic biology. *BMC Biol.* **18**, 35 (2020).
7. Silva, G. A. et al. Selective differentiation of neural progenitor cells by high-epitope density nanofibers. *Science* **303**, 1352–1355 (2004).
8. Yolamanova, M. et al. Peptide nanofibrils boost retroviral gene transfer and provide a rapid means for concentrating viruses. *Nat. Nanotechnol.* **8**, 130–136 (2013).
9. Münch, J. et al. Semen-derived amyloid fibrils drastically enhance HIV infection. *Cell* **131**, 1059–1071 (2007).
10. Kim, J. et al. In situ self-assembly for cancer therapy and imaging. *Nat. Rev. Mater.* **8**, 710–725 (2023).
11. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
12. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
13. Guo, J. et al. Cell spheroid creation by transcytotic intercellular gelation. *Nat. Nanotechnol.* **18**, 1094–1104 (2023).
14. He, P.-P. et al. Bispyrene-based self-assembled nanomaterials: in vivo self-assembly, transformation, and biomedical effects. *Acc. Chem. Res.* **52**, 367–378 (2019).
15. Gao, J., Zhan, J. & Yang, Z. Enzyme-instructed self-assembly (EISA) and hydrogelation of peptides. *Adv. Mater.* **32**, 1805798 (2020).
16. Frederix, P. W. J. M. et al. Exploring the sequence space for (tri-) peptide self-assembly to design and discover. *Nat. Chem.* **7**, 30–37 (2015).
17. Xu, T. Y. et al. Accelerating the prediction and discovery of peptide hydrogels with human-in-the-loop. *Nat. Commun.* **14**, 3880 (2023).
18. Batra, R. et al. Machine learning overcomes human bias in the discovery of self-assembling peptides. *Nat. Chem.* **14**, 1427–1435 (2022).
19. Pirtskhalava, M. et al. DBAASP v.2: an enhanced database of structure and antimicrobial/cytotoxic activity of natural and synthetic peptides. *Nucleic Acids Res.* **44**, D1104–D1112 (2016).
20. Pirtskhalava, M. et al. DBAASP v3: database of antimicrobial/ cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic Acids Res.* **49**, D288–D297 (2021).
21. Vaswani A. et al. Attention is all you need. In *Proc. 31st International Conference on Neural Information Processing Systems* (eds Guyon, I. et al.) 6000–6010 (Curran Associates, 2017).
22. McInnes, L., Healy, J. & Melville, J. UMAP: uniform manifold approximation and projection. *J. Open Source Softw.* **3**, 861 (2018).
23. Xu, Z. J. & Zhou, H. Deep frequency principle towards understanding why deeper learning is faster. In *Proc. 35th AAAI Conference on Artificial Intelligence* (eds. Leyton-Brown, K. et al.) 10541–10550 (AAAI Press, 2021).
24. Barth, A. Infrared spectroscopy of proteins. *Biochim. Biophys. Acta Bioenerg.* **1767**, 1073–1101 (2007).
25. Pavia, D. L. et al. in *Introduction to Spectroscopy,* 5th edn, 70–71 (Cengage Learning, 2015).
26. Barron, A. R. in *Chemistry of the Main Group Elements* Ch. 2.7 (Midas Green Innovations, 2014).
27. el Battioui, K. et al. In situ captured antibacterial action of membrane-incising peptide lamellae. *Nat. Commun.* **15**, 3424 (2024).
28. Marty, R. et al. Hierarchically structured microfibers of 'single stack' perylene bisimide and quaterthiophene nanowires. *ACS Nano* **7**, 8498–8508 (2013).
29. Kovacs, J. M., Mant, C. T. & Hodges, R. S. Determination of intrinsic hydrophilicity/hydrophobicity of amino acid side chains in peptides in the absence of nearest-neighbor or conformational effects. *Biopolymers* **84**, 283–297 (2006).
30. Pane, K. et al. Antimicrobial potency of cationic antimicrobial peptides can be predicted from their amino acid composition: application to the detection of 'cryptic' antimicrobial peptides. *J. Theor. Biol.* **419**, 254–265 (2017).
31. Lopetuso, L. R. et al. Commensal Clostridia: leading players in the maintenance of gut homeostasis. *Gut Pathog.* **5**, 23 (2013).
32. Zafar, H. & Saier, M. H. Gut *Bacteroides* species in health and disease. *Gut Microbes* **13**, 1848158 (2021).
33. Shi, S. H. et al. Multidrug resistant Gram-negative bacilli as predominant bacteremic pathogens in liver transplant recipients. *Transpl. Infect. Dis.* **11**, 405–412 (2009).
34. Torres, M. D. T. et al. Mining for encrypted peptide antibiotics in the human proteome. *Nat. Biomed. Eng.* **6**, 67–75 (2022).

Methods

Public AMP dataset

The public AMP sequences with lengths ranging from 6 to 15 AAs were collected from the public DBAASP database[20]. The non-AMP sequences were obtained from the UniProt databank[35] and our in-house experimental data. For the UniProt databank, we retrieved all entries satisfying the same length criterion and removed those labelled by at least one of the following keywords: antimicrobial, antibiotic, antiviral or antifungal. Moreover, we also added 33 sequences that have been validated to be non-AMPs from our previous efforts. Both datasets were filtered to exclude duplications, non-canonical AAs and terminal modifications, which resulted in 3,174 AMP sequences and 62,611 non-AMP sequences.

SAFP dataset

We collected 106 sequences composed of 6–15 AAs with C8-, C12- or C16-functionalized N termini from the DBAASP database[20]. Among the collected public AMP data, the majority of the MIC values are less than or equal to 100 μg ml^{-1} (Supplementary Fig. 27), which is set as the threshold for AMPs. According to our MIC threshold for AMPs, 93 of these peptides were active AMPs and 13 were inactive. Moreover, we chemically synthesized 204 peptides with additional N-terminal groups (C8, C12, C16, PHE, BIP, DIP, NAP, ANT, PYR, C-PRO or C-HEX) to enrich the SAFP dataset. The antimicrobial activity of synthesized SAFPs was evaluated by MIC measurements against *E. coli* and *S. aureus*. It is essential that both antimicrobially active and inactive labels be present in both the training and the validation datasets used in subsequent DL training. Of our in-house synthesized peptides, 62 were antimicrobial SAFPs and 142 were non-antimicrobial SAFPs. By concatenating the public-sourced sequences and our in-house efforts, we obtained a small dataset (310 sequences in total) with equal numbers of antimicrobial SAFPs and non-antimicrobial SAFPs.

TransSAFP pretrain module selection

We split the native peptide sequences into training, validation and test subsets at a 6:2:2 ratio in a label-stratified manner. We trained ten candidate architectures on the training and validation subsets: convolutional neural networks with one-dimensional (CNN(1D)) and two-dimensional (CNN(2D)) convolutional filters[36], recurrent neural networks with long short-term memory (LSTM)[37] or gated recurrent unit (GRU)[38] recurrent blocks, bidirectional recurrent neural networks with LSTM or GRU blocks (BiLSTM or BiGRU)[39], bidirectional recurrent neural networks with LSTM or GRU blocks and Bahdanau-style additive attention alignments (BiLSTM(A) or BiGRU(A))[40] and multi-head self-attention models with or without a transformer-style cross attention layer (MHA(T) or MHA(E))[21]. All pretraining candidates were parametrized to minimize the objective function $\mathcal{L}$:

$$\mathcal{L} = y_{\text{true}}\ \log(y_{\text{predict}}) + (1 - y_{\text{true}})\log(1 - y_{\text{predict}}) + \|x_{\text{embedding}} - x_{\text{reconstruct}}\|_2^2$$

where y_{predict} and y_{true} are the antimicrobial labels from the model prediction and the ground truth, respectively; $x_{\text{embedding}}$ and $x_{\text{reconstruct}}$ are the embedded and the reconstructed sequence features, repectively. We note that the preceding cross-entropy terms penalize the classification loss and the squared norm metric penalizes the reconstruction loss. The final pretraining architecture (MHA(T)) was selected by its competitive performance metrics on the test set and relatively efficient inference cost. The final pretraining module was obtained from retraining the selected model using an 8:2 label-stratified training–validation split on all available peptide sequences. The parameters in the pretraining module were frozen from gradient propagations in subsequent training processes.

TransSAFP transfer learning module selection

The dataset for the downstream transfer learning is merged from the public peptide and SAFP datasets. To ensure similar distributions for N-terminal modifications and antimicrobial activities in the training and the validation sets, we split the dataset at an 8:2 ratio with regard to each N-terminal type and the active/inactive antimicrobial labels. We note that at least one antimicrobially active entry and one inactive entry were included in the validation set for each N-terminal type. The downstream transfer learning module was constructed as a self-attention block appended to the latent output of the pretraining module. The sequence-immutable noise augmentation scheme is introduced as follows. We denote $X(x)$ as a state function that maps the input AA tokens (x_{AA}) of the pretraining network to the identity indices, i_{AA}, of the twenty AAs, X: $x_{\text{AA}} \to i_{\text{AA}}$. Here we implemented the X as a Euclidean distance metric, $i_{\text{AA}} = \text{argmin}\|x - x_{\text{AA}}\|_2$, which tessellates the pretraining input space into twenty Voronoi cells anchored by the discrete embeddings of AAs (x_{AA}). The sequence-immutable noise augmentation can then be defined on the noise tensor (ε) applied to each AA embedding subjected to the equality constraint: $X(x_{\text{AA}} + \varepsilon) \equiv X(x_{\text{AA}})$. To keep the noised sequence embeddings tractable, we tabulated for each x_{AA} the minimum distances to any other AA embeddings by which the uniform distribution of the noise tensors (for this AA type) was bounded. We note that such noise tensor sampling automatically satisfies the sequence-immutable constraints at negligible computational cost. The extended input of N-terminal labels was introduced by an additional embedding entry, which is concatenated to the pretraining embedding. The transfer learning module was biased during training by assigning sample weights $w_j = \frac{n(x)}{n(j)n(x_j)}$ on each sample x_j with the jth type of N-terminal modification. Here, $n(x)$ refers to the total number of sequences in the training–validation set, $n(j)$ is the total number of N-terminal types and $n(x_j)$ is the number of sequences with the jth type of N-terminal modification. The transfer learning module was trained to minimize the binary cross-entropy loss on the antimicrobially active/inactive classifications.

De novo screening of octa-SAFPs

We partitioned the entire octa-SAFP library (11×20^8 sequences) into twenty batches and distributed in parallel the screening task onto twenty Nvidia 1080 Ti graphics processing units (GPUs); the TransSAFP prediction took ~100 h in total. We retained all octa-SAFPs with prediction scores of >0.99 as potent candidates, which were then filtered for dissimilarity with the known AMP sequences on a per sequence basis. We refer to the known AMP sequences by combining the antimicrobial SAFPs verified by experimental reports and all native peptide sequences with any N-terminal modification that were predicted by the TransSAFP model to be antimicrobial (scores > 0.90). For a candidate SAFP sequence, its similarity to the known sequences is determined from the following procedure. First we aligned the candidate sequence to each of the known peptide sequences from the training dataset using the Needleman–Wunsch algorithm[41]. Then we computed the fraction of aligned residues to the length of the known sequence. After looping for all known peptide sequences, we took the maximum value of this fraction to indicate the similarity of the octa-SAFP sequences to known peptide sequences.

Computational details

All DL models were implemented using TensorFlow v.2.10 (ref. 42) on Nvidia A40 GPUs. All training sets were partitioned into learning batches of 32 entries. All models were trained using the adaptive momentum (AdaM) optimizer[43] at a constant 0.0005 learning rate through 1,000 epochs, under an early-stopping schedule that monitors and retrieves the model parameters producing the lowest validation loss. The umap-learn library was used for analysing the latent representations in the neural networks[22]. The scikit-learn library (v.1.3.0) was employed for evaluating the model performance metrics[44]. The Biopython library (v.1.80) was adopted for peptide sequence analysis[45].

Materials

AAs with an Fmoc protected group and 2-(1*H*-benzotriazole-1-yl)-1,1,3,3-tetramethyluronium hexafluorophosphate were purchased from GL Biochem Ltd. Dichloromethane was provided by Shanghai Titan Scientific. The *N*,*N*-dimethylformamide was sourced from J&K Scientific. Piperidine was provide by Sinopharm Chemical Reagent. The *N*,*N*-diisopropylethylamine was obtained from Shanghai Aladdin Biochemical Technology. The 3-(4,5-dimethyl-2-thiazolyl)-2,5-diphenyl-2*H*-tetrazolium bromide was sourced from Sanggon Biotech. The C8 and C-HEX were sourced from Shanhai Haohong Scientific. The C12 was purchased from Shaihai Yuanye Bio-Technology. The C16 was provided by Anhui Zesheng Technology. The BIP, NAP, PYR, NPN and $DiSC_3(5)$ were sourced from Shanghai Aladdin Biochemical Technology. The DIP, ANT and C-PRO were purchased from Shanghai Macklin Biochemical Co. The PHE was sourced from the Shanghai Guoyao group company. The DOPC and DOPG were sourced from AVT (Shanghai) Pharmaceutical Tech.

Bacteria strains, culture conditions and MIC assays

E. coli ATCC 25922, *S. aureus* ATCC 25923, *E. faecium* ATCC 51559, *E. faecalis* ATCC 51575 and *E. faecalis* ATCC 51299 were purchased from the ATCC. *E. coli* BNCC 186732, *E. cancerogenus* BNCC 363037, *S. epidemidis* BNCC 330867 and *A. baumannii* BNCC 254392 were provided by BNCC. *L. monocytogenes* CMCC 54004 was provided by CMCC. *S. aureus* USA 300, *S.* Typhimurium SL1344, *K. pneumoniae* BNCC 102997 and *P. aeruginosa* BNCC 360085 were obtained from our frozen stock collection. All bacteria strains were cultured in brain heart infusion broth (BHIB) at 37 °C. MIC values of bacteria strains mentioned above were determined by the microdilution method. Peptides were dissolved in PBS buffer (137 mM NaCl, 2.7 mM KCl, 8.1 mM Na_2HPO_4 and 1.9 mM KH_2PO_4) and prepared in 96-well plates by a serial twofold dilution, leading to 100 µl of solution in each well. Overnight-cultured bacterial solution was diluted and added into prepared 96-well plates, giving 200 µl in each well, with bacterial concentration of 10^6 CFU ml^{-1}. The wells without bacteria and peptides were set as positive controls, and the wells with only bacteria were employed as negative controls. Each measurement contained at least three replicates. These arrangements were applied in each plate. The plates were placed in an incubator with a setting of 37 °C. After 24 h, MIC values (the lowest concentrations that completely inhibit visible growth) were directly determined by the naked eye.

In vivo therapeutic efficacy assays

All in vivo experiments were carried out according to institutional guidelines, and corresponding experiments received approval from the Institutional Animal Care and Use Committee (IACUC) of Westlake University (IACUC animal protocol no. 22-025-WHM). The mice were kept in specific-pathogen-free conditions with a 12 h light/12 h dark cycle. The temperature was maintained between 20 and 26 °C, and the humidity was kept between 40 and 70%. Balb/c mice (female, 8 weeks old) were first subjected to 4 h food and water deprivation, after which the mice were treated with 20 mg of streptomycin (Fig. 4a). After 20 h, food and water were restricted again for 4 h followed by 100 µl of *S.* Typhimurium solution (10^9 CFU ml^{-1} in PBS) being administered to mice via oral gavage. Then, a period of two days was provided for colonization of *S.* Typhimurium in the intestines. After establishment of the infection model, each group of mice was treated with 100 µl of peptides or ciprofloxacin solutions via intraperitoneal injection with a dose of 30 mg kg^{-1}, two times, on days 3 and 4. On day 8, fresh faeces, small intestines and colons were collected from each group and homogenized in PBS buffer, after which these solutions were diluted and spread onto Brain Heart Infusion (BHI) agar plates separately in order to quantitatively measure the amount of *S.* Typhimurium. Small intestine and colon were collected and their histological statuses were analysed using H&E, PAS and AB-PAS staining. The imaging data were collected by CaseViewer (v.2.4.0).Throughout the entire assay, the body weights of mice were recorded in an Excel file every day. If the body weight of a mouse decreased by more than 20%, it was considered to have reached the humane end-point for euthanasia.

CD spectroscopy

CD measurements were performed using the CD spectrometer Chirascan V100 (Applied Photophysics). Sample solution with a certain concentration was loaded in a rectangular quartz cuvette with a 2 mm path length. CD data were collected by Pro-Data Viewer (v.4.2.0) at room temperature with a range of wavelengths from 190 to 260 nm.

Lipid membrane model

DOPC and DOPG were used to mimic bacterial inner membranes with a molar ratio of 7:3. The lipids were dissolved in chloroform and mixed with the ratio. After the evaporation of chloroform, the dry lipid mixture was dissolved by PBS buffer. Then the lipid solution was extruded 31 times through a membrane with a pore diameter of 100 nm by an Avanti Polar Lipids extruder. Finally, prepared lipid vesicle solutions were collected for other experiments.

Zeta potential measurements

The zeta potential variation of lipid vesicles upon treatment of peptides was measured by a ZetaPlus instrument (BI-200SM, Brookhaven Instrument). The data were collected by BIC Particles Solutions (v.3.6.0). A 2 ml sample solution was loaded in a cell by mixing 1 ml of vesicle solution at 1.5 mg ml^{-1} with 1 ml of peptide solution at twofold certain concentration. Each sample rested for 180 s and was measured three times to get the average values.

DLS measurements

The hydrodynamic radius variation of lipid vesicles upon treatment of peptides was measured by DLS using a wide-angle dynamic light scattering instrument (BI-200SM, Brookhaven Instrument). The data were collected by BIC Particles Solutions (v.3.6.0). The cell contained 2 ml sample solution mixed with 1 ml of vesicle solution at 1.5 mg ml^{-1} and 1 ml of peptide solution at twofold certain concentration. The cell was equilibrated for 180 s before each measurement and the hydrodynamic radius was obtained by averaging three runs.

CAC measurements

The CAC values of peptides were determined by DLS using a wide-angle dynamic light scattering instrument (BI-200SM, Brookhaven Instrument). DLS was used to measure the intensity of the peptide solution at different concentrations, and the intensity was obtained by averaging three runs. When the concentration is below the CAC, the intensity changes negligibly with increasing concentration. Once the concentration reaches the CAC, peptide monomers start to assemble and the scattering intensity of the assemblies increases exponentially as a function of concentration increase.

Membrane permeabilization assay

The NPN assay was employed to evaluate the permeability variation of the bacterial membrane. This lipophilic dye exhibits strong fluorescence when in contact with a bacteria membrane that is lipidic but shows week fluorescence under aqueous conditions. As NPN is an impermeable dye, it migrates into the membrane when the membrane is damaged (Supplementary Fig. 28). *S.* Typhimurium cells were first cultured in BHIB to an optical density at 595 nm (OD_{595}) of 0.8. The bacterial solution was centrifuged at 7,000 rpm (9,860*g*) for 5 min, and then the supernatants were removed and bacteria were resuspended by PBS buffer. This washing step was repeated three times to get rid of the culture medium. Next, 50 µl NPN solution (40 µM) was added into 50 µl of bacteria solution in a 96-well plate. Then 100 µl of peptide solution at two times a certain concentration was added into each well, while three wells that were treated with 100 µl of PBS

were set as background measurements. The fluorescence (excitation wavelength (λ_{ex}) = 350 nm, emission wavelength (λ_{em}) = 420 nm) was recorded by a plate reader (Varioskan LUX, Thermo Scientific) over time until it reached the plateau value, and the data were collected by Skanlt RE (v.7.0.2).

Resistance development assay

The development of resistance of *S.* Typhimurium against peptides or ciprofloxacin was monitored for 30 passage in series. In the beginning, the initial MICs of the peptides or ciprofloxacin against *S.* typhimurium were measured and defined as MIC_0 (initial passage). Then, the bacteria in sub-MIC wells were collected and recultured, which is the first passage. The MIC against *S.* Typhimurium (first passage) was defined as MIC_1. This process was repeated 30 times, and MIC_n/MIC_0 (n = 0–30) was calculated to investigate the development of resistance with an increasing number of passages.

Biofilm formation and eradication test

S. Typhimurium was first cultured in a flat-bottom 96-well plate for three days. During this period, the culture medium was removed and the biofilm at the bottom was washed by PBS buffer three times every 24 h to remove planktonic bacterial cells. Well-grown biofilm was treated with 200 µl of peptide solution at a certain concentration for 4 h, after which wells were washed by PBS to remove peptides and planktonic bacterial cells. Simultaneously, biofilm with treatment of PBS was set as a negative control, and empty wells were defined as a positive control. Some 100 µl of 0.1% crystal violet solution was added into each well to stain the remaining biofilm for 10 min, and then residual crystal violet solution was washed away three times by PBS buffer. Next, 95% ethanol was employed to dissolve the stain biofilm, and this process continued for 10 min. The OD_{595} was measured by a plate reader (Varioskan LUX, Thermo Scientific). Relative biofilm mass was calculated by the following function: $\text{Relative biofilm mass} = (OD - OD_{\text{positive control}})/(OD_{\text{negative control}} - OD_{\text{positive control}})$.

Confocal images of biofilm

S. Typhimurium biofilm was formed in a confocal dish (diameter of the glass bottom is 15 mm). The same steps from the biofilm eradication test were carried out, including washing and treatment of peptides or ciprofloxacin. A LIVE/DEAD BacLight bacterial viability kit was obtained from Thermo Fisher Scientific. SYTO9 stain, due to its ability to penetrate the membrane, could stain live and dead bacterial cells (green), while propidium iodide (PI), a membrane-impermeable dye, could stain only dead cells (red). The mixture of SYTO9 and PI with a ratio of 1:1 was added to treated biofilm at a working concentration of 0.3% (v/v). After 15 min of treatment, confocal images were taken by a laser scanning confocal microscope (inverted, LSM 980 with Airyscan). The imaging data were collected by Zen (v.2.3).

Morphologies of peptides by TEM

Peptides were prepared in PBS buffer with certain concentrations. Some 20 µl of peptide solution was dropped onto the front side of copper mesh. After 30 s, excess liquid was removed using filter papers. Uranyl acetate solution was then added to stain the peptides for 30 s, and excess liquid was removed using filter papers. The morphologies of peptides were examined with TEM (Talos L120C G2, Thermo Scientific) at an accelerating voltage of 120 kV.

Morphologies of peptides by cryo-EM

Peptides were prepared in PBS buffer at 1,000 µg ml^{-1} and a 3.5 µl aliquot of the solution was loaded onto glow-discharged holey carbon grids. The grids were blotted for 3 s and then plugged frozen in liquid nitrogen using Vitrobot Mark IV (Thermo Fisher Scientific) at 100% humidity and 8 °C. Cryo-EM data were then collected on a Glacios (Thermo Fisher Scientific) operating at 200 kV.

Morphologies of bacterial cells by TEM

S. Typhimurium solution was first centrifuged and washed by PBS three times and then treated with peptide solution at a certain concentration for different durations. The liquid was removed after 10 min of centrifugation (7,000 rpm/9,860*g*), and the remaining samples were treated with stationary liquid (pH 7.2, 2 vol% paraformaldehyde and 2.5 vol% glutaraldehyde) overnight. The sample was then washed by 0.1 M phosphate/cacodylate buffer (pH 7.2), three times. Subsequently, the sample was treated with osmic acid (1 wt%) and uranyl acetate (1 wt%) separately for 1 h. After every step, the sample was washed by double distilled H_2O, three times. Then, the sample was dehydrated with 30, 50, 70, 95 and 100 vol% ethanol/water and acetone with each reaction lasting for 10 min. Next, the sample was treated with a series mixture of acetone and 812 resin (2:1, 1:2) for 30 min each, and immersed in pure resin overnight. Afterward, sections of sample were prepared and stained. The morphologies of *S.* Typhimurium upon different treatments were observed by TEM (Talos L120C G2, Thermo Scientific) at an accelerating voltage of 80 kV.

Statistics and reproducibility

The method of statistical significance and the number of repetitions for the experiments (n) are described in the figure legends. For representative images (such as TEM, cryo-EM, AFM and CLSM images), experiments were performed three times independently with similar results.

Reporting summary

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

Data availability

The data that support the findings are available within the main text and the Supplementary Information and can be obtained from the corresponding authors upon request. The positive pretrained dataset was collected from the DBAASP database (https://dbaasp.org/search). The negative pretrained dataset was collected from the UniProt database (http://www.uniprot.org). Datasets and codes for the model are accessible via Science Data Bank at https://doi.org/10.57760/sciencedb.19186 (ref. 46). Source data are provided with this paper.

Code availability

The TransSAFP model can be accessed via GitHub at https://github.com/LiuHuayang27/TransSAFP (ref. 47), which is archived at the Science Data Bank at https://doi.org/10.57760/sciencedb.19186 (ref. 46).

References

35. UniProt Consortium UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic Acids Res.* **51**, D523–D531 (2023).
36. LeCun, Y. et al. Gradient-based learning applied to document recognition. *Proc. IEEE* **86**, 2278–2324 (1998).
37. Hochreiter, S. & Schmidhuber, J. Long short-term memory. *Neural Comput.* **9**, 1735–1780 (1997).
38. Cho, K. et al. On the properties of neural machine translation: encoder–decoder approaches. In *Proc. 8th Workshop on Syntax, Semantics and Structure in Statistical Translation* (eds Wu, D. et al.) 103–111 (ACL, 2014).
39. Schuster, M. & Paliwal, K. K. Bidirectional recurrent neural networks. *IEEE Trans. Signal Process.* **45**, 2673–2681 (1997).
40. Bahdanau, D., Cho, K. & Bengio, Y. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations* (ICLR, 2015).

41. Needleman, S. B. & Wunsch, C. D. A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443–453 (1970).
42. Abadi, M. et al. Tensorflow: large-scale machine learning on heterogeneous distributed systems. In *Proc. USENIX Conference on Operating Systems Design and Implementation* (eds Keeton, K. et al.) 265–283 (USENIX, 2016).
43. Kingma, D. P. & Ba, J. Adam: a method for stochastic optimization. In *3rd International Conference on Learning Representations* (ICLR, 2015).
44. Pedregosa, F. et al. Scikit-learn: machine learning in Python. *J. Mach. Learn. Res.* **12**, 2825–2830 (2011).
45. Cock, P. J. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
46. Liu, H., Song, Z., Huang, J. & Wang, H. Data archive for: "De novo design of self-assembling peptides with antimicrobial activity guided by deep-learning". *Science Data Bank* https://doi.org/10.57760/sciencedb.19186 (2024).
47. Liu, H., Song, Z., Huang, J., & Wang, H. Source codes repository for the model TransSAFP. *GitHub* https://github.com/LiuHuayang27/TransSAFP (2024).

Acknowledgements

This project was supported by the National Natural Science Foundation of China (82022038 to H.W. and 32171247 to J.H.) and the Westlake Education Foundation. This research was also supported by Zhejiang Provincial Natural Science Foundation of China under grant no. XHD23C1001. We thank the Instrumentation and Service Center for Molecular Sciences, the Instrumentation and Service Center for Physical Sciences, the Biomedical Research Core Facilities and the Supercomputer Center at Westlake University for assistance with measurements.

Author contributions

H.W. and J.H. conceptualized, supervised and founded the project. H.L., Z.S., J.H. and H.W. designed the experiments, analysed the data and wrote the paper. H.L. established the database, performed most of the experiments and participated in DL model design. Z.S. encoded the peptide sequences and designed, trained and analysed the DL model. Y.Z. and S.L. participated in the peptide synthesis, MIC experiments and in vivo therapeutic efficacy assay. B.W. and D.C. participated in the in vivo toxicity test. Z.Z. participated in the establishment of the wet-lab database. H.Z. participated in the characterization of peptides by AFM. X.F. performed the coarse-grained molecular dynamics simulations.

Competing interests

H.W., J.H., H.L. and Z.S. have filed a patent converting this work (China Patent application no. 2024112931879). The other authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41563-025-02164-3.

Correspondence and requests for materials should be addressed to Jing Huang or Huaimin Wang.

Peer review information *Nature Materials* thanks Cesar de la Fuente, Sonia Henriques and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Huaimin Wang, Jing Huang

Last updated by author(s): Jan 3, 2025

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Agilent Open Lab Control Panel (LCMS-online) was used to collect the liquid chromatography mass spectrometry; Zen (2.3) was employed for imaging data collection; Skanlt RE (7.0.2) was used to collect the data of bacterial growth curves, hemolysis, MMT, MIC, DiSC3(5), NPN measurements; BIC Particle Solutions (3.6.0) was employed for DLS and Zeta potential data collection; Pro-Data Viewer (4.2.0) was used to collect CD data; Velox (3.8.0) was used to collect TEM imaging data; Excel 2019 was used to record and collect the data, including mice body weight, villi height, and bacterial burden; CaseViewer (2.4.0) for H&E, PAS, and AB-PAS imaging data collection. OMNIC (9.7.46) was used to collect FTIR data;
Data analysis	TensorFlow (2.10) was used to construct the deep-learning models and Scikit-learn (1.3.0) was employed for machine learning models; The preparation of the CGMD systems used the CHARMM (c42b2), Auto-Martini (0.2.0), VerMoUTH-Martinize (0.9.1), amd Packmol (20.11.1) packages. The visualization of the simulation boxes used UCSF ChimeraX (1.7); BioPython (1.80) was used to analyze the similarity of peptide sequences; Excel 2019, Origin 2018 and Graphpad Prism 8 were used to analyze and plot data; Pro-Data Viewer (4.2.0) was employed for DLS data analysis; The TransSAFP model can be accessed at https://github.com/LiuHuayang27/TransSAFP.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All the data that support the findings are available within the main text and the supplementary materials. Positive pre-trained dataset was collected from DBAASP database (https://dbaasp.org/search). Negative pre-trained dataset was collected from UniProt database (http://www.uniprot.org). Datasets and codes for the model are accessible at the following DOI: https://doi.org/10.57760/sciencedb.19186.

Research involving human participants, their data, or biological material

Policy information about studies with human participants or human data. See also policy information about sex, gender (identity/presentation), and sexual orientation and race, ethnicity and racism.

Reporting on sex and gender	N/A
Reporting on race, ethnicity, or other socially relevant groupings	N/A
Population characteristics	N/A
Recruitment	N/A
Ethics oversight	N/A

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample sizes were determined according to previous literature on similar measurements (triplicate at least to avoid random error). The sample sizes for the in vivo experiments are listed as below: For the in vivo toxicity evaluation, n = 5; For the S. typhimurium infection model, n = 5; The sample size was estimated based on our pilot experiments. We found that five mice are sufficient to observe significant differences between control and experimental groups while avoiding unnecessary wastage of resources.
Data exclusions	No data were excluded
Replication	All experiments were replicated at least three independent tests, and all attempts at replications were successful.
Randomization	All samples and mice were randomly allocated into experimental groups.
Blinding	Data acquisition and analysis were performed by investigators blinded to the groups.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☐	☒ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern
☒	☐ Plants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	GES-1 was obtained from Hunan Fenghui Biotechnology Co., Ltd (China).
Authentication	None of the cell lines used were authenticated.
Mycoplasma contamination	No mycoplasma contamination was found in the used cell lines.
Commonly misidentified lines (See ICLAC register)	No misidentified cell lines were used.

Animals and other research organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research, and Sex and Gender in Research

Laboratory animals	Balb/c mice (female, 6-8 weeks) were obtained from the laboratory animal resources center (LARC) at Westlake University.
Wild animals	The study did not involve wild animals.
Reporting on sex	N/A
Field-collected samples	The study did not involve samples collected from the field
Ethics oversight	All mice were handled in accordance with institutional guidelines, and all animal experiments were approved by the Institutional Animal Care and Use Committee (IACUC) of Westlake University (IACUC Protocol #22-025-WHM).

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Plants

Seed stocks	*Report on the source of all seed stocks or other plant material used. If applicable, state the seed stock centre and catalogue number. If plant specimens were collected from the field, describe the collection location, date and sampling procedures.*
Novel plant genotypes	*Describe the methods by which all novel plant genotypes were produced. This includes those generated by transgenic approaches, gene editing, chemical/radiation-based mutagenesis and hybridization. For transgenic lines, describe the transformation method, the number of independent lines analyzed and the generation upon which experiments were performed. For gene-edited lines, describe the editor used, the endogenous sequence targeted for editing, the targeting guide RNA sequence (if applicable) and how the editor was applied.*
Authentication	*Describe any authentication procedures for each seed stock used or novel genotype generated. Describe any experiments used to assess the effect of a mutation and, where applicable, how potential secondary effects (e.g. second site T-DNA insertions, mosiacism, off-target gene editing) were examined.*