nature biotechnology

ARTICLES

https://doi.org/10.1038/s41587-022-01226-0

Check for updates

Identification of antimicrobial peptides from the human gut microbiome using deep learning

Yue Ma[1,2,5], Zhengyan Guo[3,4,5], Binbin Xia[1,2,5], Yuwei Zhang[2,3,5], Xiaolin Liu[1,2], Ying Yu[1,2], Na Tang[2,3], Xiaomei Tong[1], Min Wang[2,3], Xin Ye[1,2], Jie Feng[2,3], Yihua Chen[2,3] ✉ and Jun Wang[1,2] ✉

The human gut microbiome encodes a large variety of antimicrobial peptides (AMPs), but the short lengths of AMPs pose a challenge for computational prediction. Here we combined multiple natural language processing neural network models, including LSTM, Attention and BERT, to form a unified pipeline for candidate AMP identification from human gut microbiome data. Of 2,349 sequences identified as candidate AMPs, 216 were chemically synthesized, with 181 showing antimicrobial activity (a positive rate of >83%). Most of these peptides have less than 40% sequence homology to AMPs in the training set. Further characterization of the 11 most potent AMPs showed high efficacy against antibiotic-resistant, Gram-negative pathogens and demonstrated significant efficacy in lowering bacterial load by more than tenfold against a mouse model of bacterial lung infection. Our study showcases the potential of machine learning approaches for mining functional peptides from metagenome data and accelerating the discovery of promising AMP candidate molecules for in-depth investigations.

The emergence and quick spread of antibiotic-resistant pathogens causes increasing numbers of difficult-to-treat infections and poses a threat to global health, as infection-related fatalities caused by drug-resistant pathogens are predicted to account for the highest number of deaths around the world by 2050 (ref. [1]). In 2017, the World Health Organization published a priority pathogens list for new antimicrobial drugs, collectively named as ESKAPE[2]. Among these pathogens, Gram-negative bacteria, such as carbapenem-resistant Enterobacteriaceae (CRE), are of particular concern owing to their ability to rapidly develop antibiotic resistance[3]. However, a combination of lack of economic incentives and market failure has led to a gradual decline in discovery and development efforts and very few antibiotics being commercialized in recent decades[4].

A large number of existing antibiotics and many other medicines originate from microbial metabolites. Among the bioactive secondary metabolites, short peptides attract extensive attention for their high diversity and wide bioactivity spectra, and the particularly large group of antimicrobial peptides (AMPs) from bacteria have been used for treating bacterial, fungal and viral infections and even cancer[5]. Antibiotics have already been developed from bacterial AMPs, primarily from non-ribosomally synthesized peptides and ribosomally synthesized and post-translationally modified peptides. In addition, class II and class III bacteriocins[6] can be ribosomally synthesized and function without modification. This means that they can be directly identified from microbial genomes, similarly to AMPs from eukaryotic genomes, such as human LL37 (cathelicidin)[7]. AMPs differ from small-molecule antibiotics as they have lower susceptibility to resistance development in pathogens[8] and stronger phylogenetic barriers within bacteria against horizontal transfer of developed resistance[9].

Developments in sequencing technology have enabled the in-depth understanding of microbiomes, especially in the human gut, the contributions of which to host metabolic and immune health increasingly are being recognized[10–12]. The gut microbiome encodes highly diverse genes, being one of the largest reservoirs for antibiotic-resistant genes[13]. At the same time, as a result of long-term competition and co-evolution, it is expected to produce a large number of antimicrobials against even multi-drug-resistant (MDR) bacteria[14]. Multiple cases have demonstrated that, in the human gut, AMPs are capable of modulating inter-species competition and maintaining community structure. For instance, a gut microbial lantibiotic peptide restored host resistance to vancomycin-resistant *Enterococcus*[15], with additional reports of their effects on host immunity[16]. According to bioinformatics analysis, a large number of potential AMP families in the human gut microbiome remain to be studied in depth[14].

The large number of potential AMPs derived from the human gut microbiome, thus, in theory, could serve as a source of candidates against infectious bacteria[17]. Until now, however, the discovery of AMPs has remained largely experimentally driven, and bioinformatic approaches have remained challenging owing to the relatively short length and low sequence similarities among AMPs[18].

Artificial intelligence approaches, in particular natural language processing (NLP) methods, can learn sequence features autonomously and might enable the identification of candidate AMPs by identifying features from genome sequences, even short sequences with low homology. Machine learning has already successfully identified small molecules with antibiotic effects, such as halicin[19], as well as inhibitors for DDR1 enzyme from small molecules[20]. In addition, short AMPs have been recently generated in silico through the combination of deep learning and physiochemical selection from controlled data generation[21–25], establishing the feasibility of such an approach.

Here we demonstrate that the combination of neural network models (NNMs) for autonomous learning of AMP sequence features and large-scale human microbiome data resources can discover AMPs with high antibacterial potency (Fig. 1). We constructed

[1]CAS Key Laboratory of Pathogenic Microbiology and Immunology, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China. [2]University of Chinese Academy of Sciences, Beijing, China. [3]State Key Laboratory of Microbial Resources, Institute of Microbiology, Chinese Academy of Sciences, Beijing, China. [4]Institute of Medicinal Biotechnology, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China. [5]These authors contributed equally: Yue Ma, Zhengyan Guo, Binbin Xia, Yuwei Zhang. ✉e-mail: chenyihua@im.ac.cn; junwang@im.ac.cn

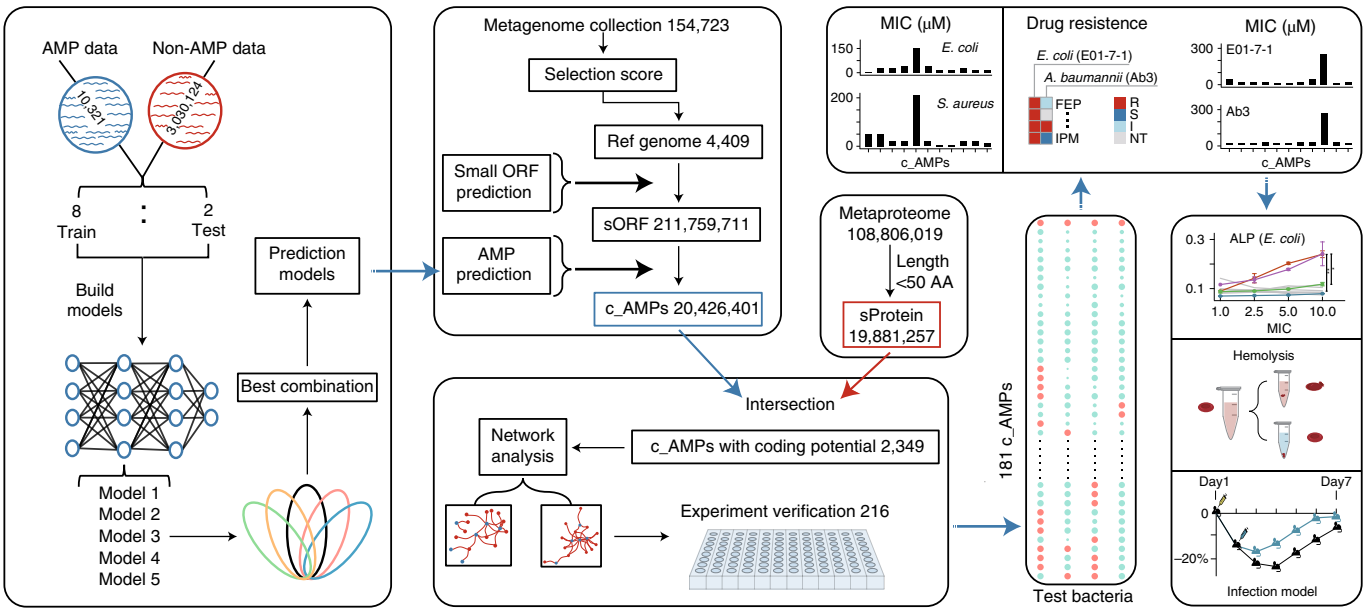


Fig. 1 | Schematic representation of study workflow. In this study, we started from collecting sequences to build training and test sets and then built and optimized NNMs to form the AMP prediction pipeline (left). We then mined metagenomic and metaproteomic data for potential AMPs, further filtering using correlational network analysis between candidate AMPs and bacteria, resulting in candidate AMPs for chemical synthesis and in vitro validations (middle). Promising candidates are selected from initial screening and further subjected to efficacy tests against MDR bacteria, in vivo experiments in an animal model of a bacterial lung infection and mechanistic assays (right).

and combined multiple NNMs and mined potential AMPs in large metagenomic data. In total, 216 novel peptides were chemically synthesized, and at least 181 of them were confirmed with antibacterial activity (83.8%). Further selection identified AMPs with high efficacy against MDR, Gram-negative bacteria and potency against in vivo infection in an animal model. Our work highlights the potential of the combination of machine learning and large metagenomic datasets to improve AMP prediction and identify new classes of functional AMP molecules.

Results

Combining NLP models to create a unified pipeline for AMP identification. We applied NLP algorithms to construct AMP prediction models, including five NNMs belonging to three classes (Fig. 2b). Our basal model was based on Veltri et al.,[26] an NNM with long short-term memory (LSTM) as the core layer that has already been shown as effective for AMP identification. A second model replaced the LSTM layer with an Attention layer, resulting in an ATT model. Lastly, we included a transformer-based Bidirectional Encoder Representations from Transformers (BERT) model[27]. Hyperparameters of the NLP models were tested and screened using independent datasets (Methods), and all models converged rapidly during training (Extended Data Fig. 1b).

We optimized the performance of the models and formed a unified pipeline for AMP identification. For ATT and LSTM, we have also included balanced training datasets (named b_ATT and b_LSTM, respectively) for comparison and later re-trained the two models using the full non-AMP training dataset. Initially, our key evaluation parameter Precision (defined as the proportion of true positives within all predicted positives) was less than 50% for each model alone (Supplementary Table 1). Re-analysis separating sequences by length revealed that precisions were higher for peptides ≤50 amino acids (AAs) (55% of previously confirmed AMPs were shorter than 50 AAs (Supplementary Table 1)), so we retained smaller peptides in the test set for the re-evaluation of the models.

We found that the proportion of true-positive (TP) and false-positive (FP) sequences identified by different models varied greatly: LSTM predicted 56 times FPs and 11% more TPs than BERT, and ATT predicted 46% more FPs and 5% more TPs than BERT (Supplementary Table 1). These differences do not simply reflect the distinct sensitivities of the different models, because, when we compared the TPs and FPs shared by all models (TPs: 1,678 and FPs: 3,981) and extracted the prediction vectors of these TPs and FPs in the last hidden layer of each model, only less than 0.3% of all pairwise correlations of these sequences among five models were significantly correlated (false discovery rate (FDR) < 0.05) (Fig. 2c and Supplementary Table 2).

As their prediction biases were independent of each other, we explored combining these different models to further improve Precision and to decrease FPs. We eventually tested the intersection of various combinations of models (2–5 models) and evaluated the combinations of models using Precision, Recall and Area Under the Precision Recall Curve (AUPRC; Methods). Ranking by AUPRC, as well as these two parameters, revealed that the combination with the highest Precision was a combination of three models at 91.31% (ATT, LSTM and BERT, with a ~15% improvement compared to the best performance of a single BERT model), with Recall reaching 83.32% and highest AUPRC of 0.9244 (Fig. 2d and Supplementary Table 3). A comparison with other currently available AMP prediction methods using the same test dataset revealed that our pipeline surpassed all others in terms of AUPRC (Fig. 2e) and Precision (other tools: 27.21~2.67%), and, whereas four of the tools predicted 0.66–6.75% more TP AMPs than ours, the FPs were 74–~182-fold higher than our pipeline (Table 1). These results support that our pipeline combining multiple NLP models is a robust approach for AMP identification from sequence data.

Identification of many candidate AMPs in large metagenomic cohorts. The human microbiome has a complex community structure, and it is possible that the constituent microbes could employ many AMPs to help compete for resources or stabilize the

community structure. We, therefore, conducted AMP mining in human microbiome data. The identification of small open-reading frames (sORFs; 5–50 AAs in length) in metagenomic data is extremely time- and computational resource-consuming (need to assembly); we, therefore, used direct sORF prediction in representative genomes that were previously assembled based on metagenomics studies[28]. A total of 154,723 representative genomes of organisms present in human microbiome samples were filtered based on more than 90% completeness (47.7%; Fig. 3a) and selection score (Methods). We eventually focused on 4,409 qualified representative genomes (Fig. 3b).

Sequential prediction and filtering of AMPs were performed to refine the putative AMP list. In total, 20,426,401 were predicted as putative AMPs from non-redundant sORF sequences contained within the representative genomes. Given that in silico prediction of sORFs does not ensure that they are expressed as proteins, we subsequently cross-checked with available metaproteomic data (19,881,257 non-redundant peptides) to screen for 'very likely' expressed peptides and eventually identified 2,349 as candidate AMPs (c_AMPs) with a length distribution ranging from 6 AAs to 50 AAs (Extended Data Fig. 1c).

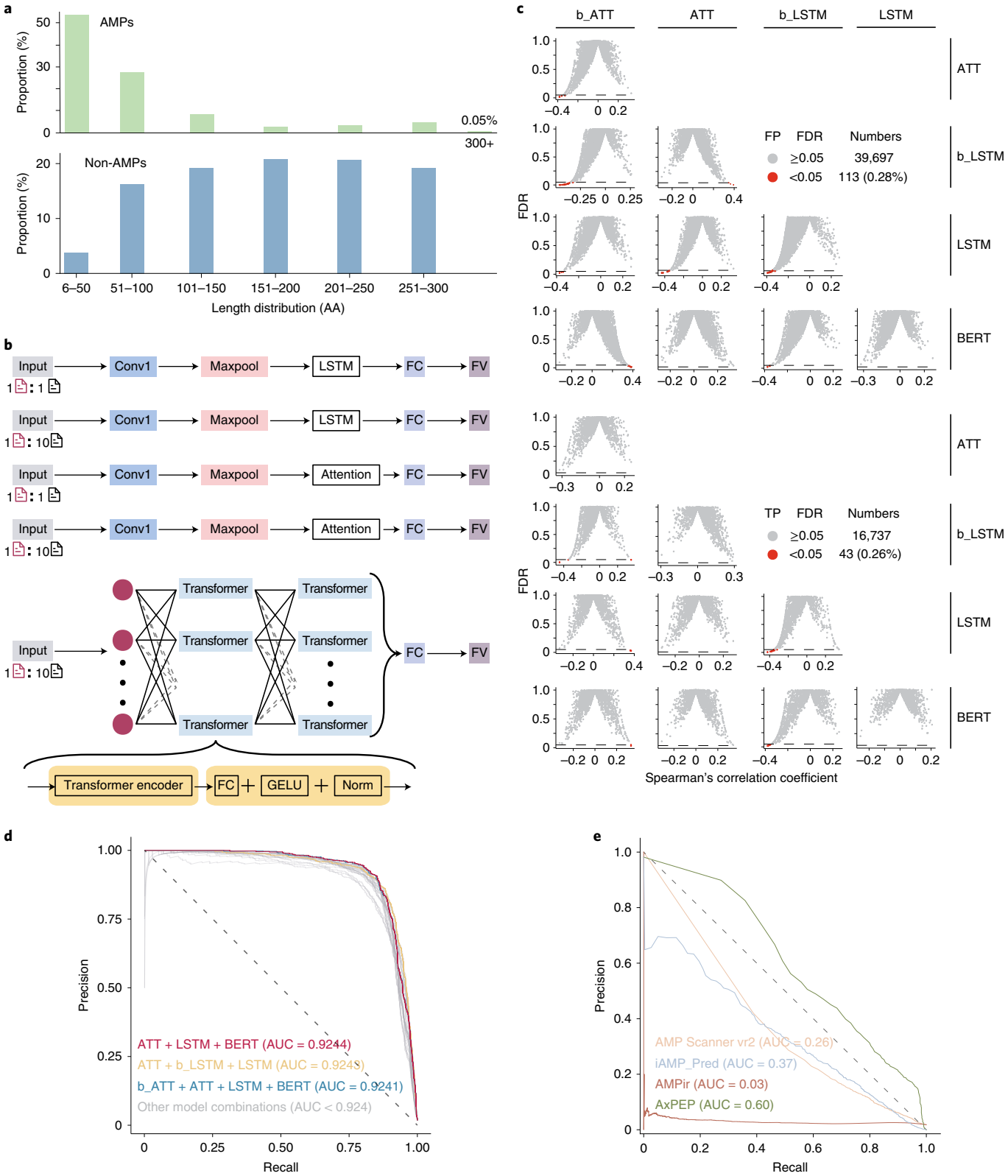
We subsequently analyzed the c_AMPs based on an association network analysis using metagenomic datasets from large cohorts. Given the known effects of some AMPs in regulating and stabilizing community structure[29,30], we hypothesized that c_AMPs with strong negative correlations with members of a microbiome thus potentially inhibit bacterial growth and are more likely to be functional; and this network could help further eliminate FPs in our discovery. We accordingly examined metagenomic datasets totaling 11,011 samples from 15 independent cohorts, most with more than 100 samples (Fig. 3c). For each cohort, we used a mapping-based approach to calculate the c_AMP abundances, as well as bacterial relative abundances at both the genus and species level, to construct c_AMP–microbial correlation networks, which were different from bacterial correlations because one potential AMP could be encoded by multiple bacteria and vice versa. By retaining those negative correlations having an FDR ≤ 0.05 and appearing in seven or more cohorts, we obtained 368 and 266 non-redundant c_AMPs, respectively, for species-level and genus-level networks, with 241 shared (Fig. 3d,e).

Experimental validation verifies high proportion of functional AMPs. We then chemically synthesized the 241 shared c_AMPs for verification and characterization. Twenty-five peptides failed after three rounds of chemical syntheses and resulted in 216 c_AMPs. We then examined the antibacterial activity of the 216 c_AMPs against *Staphylococcus aureus* (American Type Culture Collection (ATCC) 6538), *Bacillus subtilis* (ATCC 23857), *Escherichia coli* DH5α and *Pseudomonas aeruginosa* (ATCC 15692) (two Gram-positive and two Gram-negative) at 60 μM concentration in liquid media. This identified 181 peptides with antibacterial activity (that is, inhibiting the growth of at least one bacteria), representing a positive rate of 83.8% (181/216). Simultaneously, a negative set of ten peptides was randomly chosen from a predicted non-AMP set, of which six were successfully synthesized, among which only one is effective (minimal inhibitory concentration (MIC) ≤ 125 μM), indicating that our method has a low FN rate of approximately 16.67% (Supplementary Table 4), close to the calculated FN rate (1-Recall; Supplementary Table 3). The 11 c_AMPs with the highest antibacterial activity were selected with criteria that the occurrences of c_AMPs were among the top ten c_AMPs against each strain (Fig. 4a and Supplementary Tables 5 and 6). The remainders are shown in Extended Data Fig. 2; potential origins of the 181 AMPs are provided in Supplementary Table 7. Among those 11 potent c_AMPs, seven were originally found in the species genome from *Bacteroides*, which is known to contain potential probiotic species, suggesting that the AMPs might participate in pathogen inhibition. A total of 157 c_AMPs showed inhibitory effects against Gram-negative bacteria. Retrospectively, we compared some available AMP prediction tools to examine our 181 verified c_AMPs and found that all the tools predicted 30–178 of the c_AMPs to be TPs (Supplementary Table 8).

We then evaluated the similarity of the 181 c_AMPs with previously reported antibacterial AMPs. The results support that our method could identify AMPs based on internal relationships of AAs in the sequences instead of the sequence similarity with AMPs used in the training dataset. When our training dataset involved AMPs of microbial and eukaryotic origin without dissecting different AMP classes, the highest identity among our verified c_AMPs to the closest AMPs used in the training dataset was only 61.4%, and most had less than 40% identity (Fig. 4b). When we calculated the identities among our c_AMPs and non-AMPs of the training dataset, the closest identities between AMPs and non-AMPs had a median of 33.3%, higher than that of our c_AMPs versus known AMPs (median of 31.1%) (Extended Data Fig. 3a). In addition, our verified c_AMPs were not merely predicted based on AA composition, as they have seemingly different AA compositions in comparison to known AMPs (Extended Data Fig. 3c), displaying a relatively higher proportion of Glu, Lys, Asn and Gln residues (1.25-, 1,29-, 1.35- and 1.64-fold, respectively) and fewer Cys, His, Leu and Trp (0.56-, 0.72-, 0.68- and 0.61-fold, respectively); details are provided in Supplementary Table 9.

Shortlisting the most potent peptides against Gram-negative bacteria. We examined the efficacy of the top 11 c_AMPs from our initial antibacterial activity screening against the commonly studied antibiotic-resistant, Gram-negative bacterial pathogens. All 11 c_AMPs showed inhibitory activities against *E. coli* DH5α and *P. aeruginosa* (ATCC 15692) (Fig. 4a). Then, the species of bacteria for antibacterial testing were expanded to *Acinetobacter baumannii* (ATCC 19606), *Klebsiella pneumoniae* (NCTC 5056) and *Enterobacter cloacae* (ATCC 13047) and performed quantification of MIC. For *K. pneumoniae*, a total of ten c_AMPs have MIC < 25 μM, and, for other Gram-negative species, including *A. baumanii*, *E. cloacae* and *E. coli*, at least one c_AMP reached MIC of 25 μM. Also, a total of nine c_AMPs reached <20 μM in their MIC against at least one Gram-negative bacterial species (Fig. 4c

Fig. 2 | Establishing AMP prediction pipeline combining NLP models. a, Length distribution of AMPs and non-AMPs initially collected from various databases for training. Most AMPs were below 300 AAs. Later, we selected length-matched sets of AMPs and non-AMPs for training (Extended Data Fig. 1a). **b**, Summary of five NLP models included for testing and building the prediction pipeline, including LSTM models using two training datasets (LSTM using ten times of non-AMPs than AMPs for training and b_LSTM using equal amounts of AMPs and non-AMPs for training (Methods)); Attention models using two training datasets (ATT and b_ATT; training data ratios are identical to LSTM and b_LSTM (Methods)); and one BERT model. FC, fully connected; FV, feature vector; GELU, Gaussian error linear unit (the activation functions). **c**, Overview of correlations between vectors of prediction values in the last hidden layer of each NNM, for all shared FPs (up) and TPs (down). In each panel, the FDRs of Spearman correlations of prediction vectors between the prediction value vectors for each sequence are plotted on the *y* axis, and significant correlations are marked in red. The *x* axis indicates the Spearman correlation coefficients. Overall, less than 0.3% of the correlations were significant. **d**, AUPRCs for different combinations of NLP models. The three with highest values are shown, and the combination of ATT + LSTM + BERT had the highest AUPRC (0.9244) as well as the highest Precision (91.31%). **e**, AUPRCs for representative non-NLP methods that we used for comparison are included in Table 1.



and Supplementary Table 10). Compared to known AMPs with MIC measures in all of our collected AMP databases (1,009 MIC values for *E. coli* in total, 858 for *S. aureus*), our c_AMPs had MICs among the lowest (~1% percentile) of all AMPs. In fact, only seven previously confirmed AMPs have lower MICs against *E. coli* than our selected c_AMPs (Fig. 4d and Supplementary Table 11). Among those, Beta defensin 2, Cathelicidin BF, Defensin_NP-1, Magainin 2 and Mastoparan-like peptide 12c precursor were also synthesized and used as positive controls in our study, and their MICs are overall in the same magnitude as reported in previous research (Supplementary Table 12).

MDR, Gram-negative bacteria, such as CRE and members of ESKAPE (*K. pneumoniae, A. baumannii, P. aeruginosa and Enterobacter spp*), are of particular concern[31]. To assess the abil-

Table 1 | Comparison of performance between our pipeline and other currently available AMP prediction tools

Method	TP	FP	TN	FN	Precision (%)	F1 score (%)	MCC (%)
Ours	904	86	66,924	181	91.31	87.13	87.03
Macrel	785	2,100	64,910	300	27.21	39.55	43.03
AxPEP	1,040	6,411	60,599	45	13.96	24.37	34.61
iAMP-2L	910	7,044	59,966	175	11.44	20.13	28.60
iAMP Pred	933	9,833	57,177	152	8.67	15.75	24.48
AMP Scanner v2	965	15,736	51,274	120	5.78	10.85	19.05
AMPir	264	7,687	59,323	821	3.32	5.84	5.01
iAMP-CA2L	337	12,303	54,707	748	2.67	4.91	4.09

We calculated the incidences of TP (true positive), FP (false positive), TN (true negative) and FN (false negative) predictions of each tool and then Precision, F1 score and Matthews correlation coefficient (MCC). Our pipeline has the highest Precision, and seven tools (Macrel[60], AxPEP[61], iAMP-2L[62], iAMP-Pred[63], AMP-scanner version 2 (ref. [26]), AMPir[64] and iAMP-CA2L[65]) have higher FP rates despite predicting more TPs.

ity of the selected c_AMPs against MDR, Gram-negative bacteria, we performed growth inhibition assays against ten clinical isolates of MDR *A. baumannii* (three strains), *E. coli* (three strains)[32] and *K. pneumoniae* (four strains with the antibiotic resistance profiles displayed in Fig. 4e). In brief, all of the assayed strains are known to be resistant to third-generation cephalosporins ceftazidime, ceftriaxone, cefepime and sulbactam cefoperazone (CAZ, CRO, FEP and SCF); all of the *K. pneumoniae* and *E. coli* clinical isolates, as well as *A. baumannii* Ab8, were resistant to at least one carbapenem antibiotic ertapenem, imipenem or meropenem (ETP, IPM or MEM). In our assay, c_AMP1043 reached <10 μM MIC against all of the clinical isolates, and seven c_AMPs reached <20 μM MIC against at least nine clinical isolates (Fig. 4f and Supplementary Table 13). Therefore, our selected candidates have low similarity to known AMPs but have broad-spectrum and potent antibacterial activity, including against MDR, Gram-negative bacteria.

Selected c_AMPs are effective against bacterial lung infections in a mouse model. Before conducting in vivo infection experiments with mice, we evaluated the toxicity of the 11 c_AMPs against eukaryotic cells, including HCT116 cells (human colorectal cancer cell line) and fresh human erythrocytes. We performed hemolysis and cytotoxicity assays on these 11 peptides using various concentrations and estimated respective IC50/CC50 values (Fig. 5a and Supplementary Table 14). Combining these results and our MIC data for the c_AMPs against MDR *K. pneumoniae* (ATCC 700603) (Supplementary Table 15 and Fig. 4c), we selected c_AMP1043, c_AMP593 and c_AMP575 for in vivo analysis using a mouse model infected with *K. pneumoniae* for which we monitored body weight recovery as the primary readout[33,34]. Compared to the vehicle-treated control group, infected mice treated with the c_AMPs showed significantly faster rates of body weight recovery (Fig. 5c); additional colony-forming units (c.f.u.) and real-time polymerase chain reaction assays further confirmed a significant decrease of *K. pneumoniae* load in mouse lung 24 h after c_AMP treatments (Fig. 5e), indicating that c_AMPs reduced severity of the bacterial infection. Specifically, whereas roughly half of the vehicle-treated control group displayed reduced body weight for more than 7 d, all the mice treated with c_AMPs had recovered to their original body weight by this point. The results demonstrate that the three c_AMPs exert antibacterial activity against lung infection without any obvious adverse effects on the host and deserve further investigations.

Mechanisms of action of c_AMPs. We then focused on elucidating the mechanisms of action for selected AMPs, particularly c_AMP1043, which showed the highest efficacy in vitro against *K. pneumoniae*. The common mechanism of AMPs is bacterial lysis by forming pores on cell walls or membranes[35–38], and our time–kill assays asserted that the bactericidal effects were manifested within 2–6 h (Fig. 5b). We first used transmission electron microscopy (TEM) to monitor potential morphological changes in *E. coli* DH5α cells upon c_AMP1043 treatment (MIC = 2 μM, treated with 1× and 10× MIC for 5 h) and dose-dependent wrinkling exhibited in the AMP-treated cells, with cell content leaking out at the 1× MIC dose; at higher concentration, the cells collapsed and lysed (Fig. 5d and Extended Data Fig. 4a,b), suggesting that the integrity of the cell wall decreased. We carried out four experiments to determine the exact cellular components affected by c_AMP1043. Whereas measuring the extracellular levels of alkaline phosphatase (ALP) from collapsed *E. coli* DH5α cells (treated with 1×, 2.5×, 5× and 10× MIC for 1 h—the four c_AMP concentrations used in all following mechanism assays) revealed no change compared to non-treated bacteria (Fig. 5f, ALP), we did detect increased fluorescence intensity with the membrane-impermeable probe propidium iodide (PI). After treating the cells with c_AMPs for 1 h, we found that the PI intensity increased as the c_AMP1043 concentration increased, thus indicating cell membrane disruption (Fig. 5f, PI). We next applied the fluorescent probe 1-*N*-phenylnaphthylamine (NPN) to examine the permeability of the outer membrane of *E. coli* DH5α cells (treated with c_AMPs for 30 min). The dose-dependent fluorescence intensity enhancement that we observed suggested that c_AMP1043 damaged the integrity of the outer membrane (Fig. 5f, NPN). Additionally, we employed the potentiometric fluoroprobe 3,3′-dipropylthiadicarbocyanine iodide ($DiSC_3(5)$). We detected a c_AMP1043 dose-dependent increase in $DiSC_3(5)$ signal intensity after cells were treated with c_AMPs for 1 h (Fig. 5f, $DiSC_3(5)$), findings demonstrating that c_AMP1043 can disrupt the membrane potential. These results establish that the disruption of the outer membrane potential contributes to the observed inhibitory effects of c_AMP1043 against the growth of Gram-negative bacteria.

The potential membrane disruption and cell wall damage-related effects of the other ten c_AMPs were also assessed in assays against *E. coli*. Using the aforementioned ALP, PI, NPN and $DiSC_3(5)$ assays, we identified eight peptides acting on membranes and causing disruption, among which two also lysed cell walls (Extended Data Fig. 4c). Still, two remaining peptides, c_AMP593 and c_AMP575, which were effective in treating mouse infection models, and our current combination of functional assays did not result in potential mechanisms. Notably, after 30 d of continuous treatment with the c_AMP1043, no resistance could be detected in *E. coli* DH5α (Extended Data Fig. 4d). Because we included AMPs in our training datasets without distinguishing their chemical classes or mechanisms, aiming to discover AMPs via hidden sequence features, the results here indicate that our pipeline might also capture AMPs with different functional mechanisms.

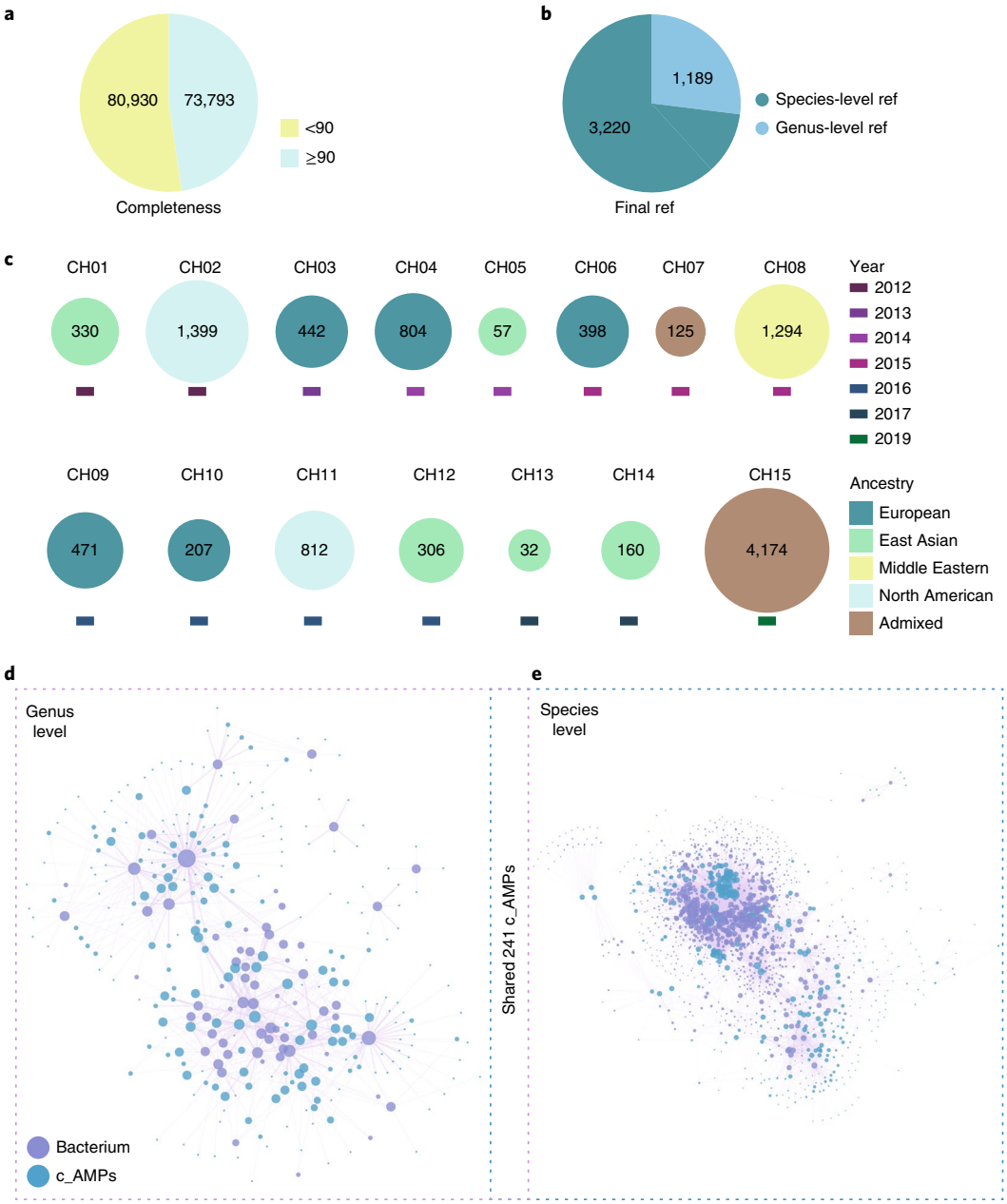


Fig. 3 | Mining candidate AMPs from metagenomic data. a, Overview of representative genomes and their completeness that were assembled from large-scale metagenomic data, from the study of Pasolli et al.[28] (Methods). **b**, Overview taxonomical levels that can be assigned for the selected representative genomes, after we filtered on contig numbers, completeness and level of ornamentations (Methods). **c**, Summary of cohorts whose metagenomic data were used for AMP mining, with year of publication and cohort size and ancestry information provided for each cohort. **d**,**e**, Network of negatively correlated AMPs and bacterial genera (**d**) and species (**e**), respectively, showing only significant correlations (Spearman correlation and FDR < 0.05) appearing in more than six cohorts, which we used for further selecting AMPs with high occurrences and potential effectiveness (Methods). Light purple dots represent AMPs; light blue dots represent bacteria.

Discussion

The recent surge in applying artificial intelligence approaches in biomedical research has brought up many new opportunities, including image-processing-related tools for clinical diagnostics of eye[39], lung[40], cancer tissue pathologies[41] and antibiotics[19,42]. In this study, by treating peptide sequences as text and applying NLP methods, together with using large datasets of gut microbiome sequences, we identified many microbial functional peptides. Previously, typical approaches for identifying proteins with similar functions were based on sequence alignment, such as BLAST, or identifying conserved motifs and domains using hidden Markov models[14,43,44]; however, it is difficult to apply these approaches with shorter peptides, particularly for those lacking significant homology[45]. In the training dataset, we observed significantly higher homologies among known AMPs (median, 88.9%) than between known AMPs and non-AMPs (a median of 28.6% and 33.3% for the balanced and unbalanced datasets, respectively, both $P < 0.05$; Extended Data Fig. 3b). In contrast, NLP methods based on different NNMs have been shown to be more effective in identifying particular subsets of sequences; in our example, validated c_AMPs share less than 50% of homology to known AMPs.

We used three different NLP models and noted that each had independent predictive biases; we, therefore, optimized predictive

performance by combining these models into a unified pipeline. By considering the intersection of three models, we obtained >0.92 AUPRC and >91% Precision in distinguishing AMPs from non-AMPs in the testing datasets. It is worth noting that ever-accumulating AMPs in reference databases, improvement on data quality and methodological advances in NLP models would further improve the power of mining AMPs. Although our pipeline used a relatively straightforward combination of models (prediction score >0.5 in each model) to achieve a balance between Precision and Recall, further optimization on various cutoffs for prediction scores, as well as inclusion of both NLP and non-NLP methods, can be carried out in future applications.

We applied our AMP-mining methods to human gut microbiome data, which have been anticipated to be a large reservoir of functional proteins, including AMPs[46,47]. For context, although microbial AMPs have a long history of experimental discovery[48], and their medical potentials have been widely recognized, it is clear that the verified AMPs account for only a small fraction of all the AMPs in nature[5]. Our approach theoretically enables further mining of AMPs from metagenomic sequences and can, therefore, provide candidates for various applications (including genetic engineering). We cross-checked with metaproteomics data to ensure the expression of potential peptides and further selected AMPs that were significantly negatively correlated to bacterial species in multiple metagenomic cohorts[29,49], to select potential c_AMPs for follow-up functional studies. With potentially over-stringent inclusion criteria, we still identified more than 200 candidates, and 181 were confirmed to exert antimicrobial activities. Given that some chemically synthesized peptides might lose certain biological functions by not generating secondary structures, and only four bacterial strains were used in the initial antibacterial activity assays, the inactive c_AMPs that were filtered out might also exert antimicrobial properties for other bacteria.

Follow-up investigations confirmed that our method can discover diverse, unreported AMPs, which exert distinct antimicrobial activity spectra and employ different mechanisms of antibacterial action. Several studies with manual filtering of features for AMP prediction or optimization were carried out previously[50–53]. Our pipeline also did not particularly rely on sequence similarity or AA composition for AMP prediction (Fig. 4b and Extended Data Fig. 3c) because our verified c_AMPs had low homology with known AMPs, and mechanistic analysis of selected c_AMPs confirmed that our approach did not selectively identify AMPs of particular action mechanism (Extended Data Fig. 4c). Indeed, in our training dataset, we used solely the sequence information of experimentally verified AMPs, and, yet, our approach managed to detect deep, hidden features within them and consequently discovered diverse AMPs in metagenomes.

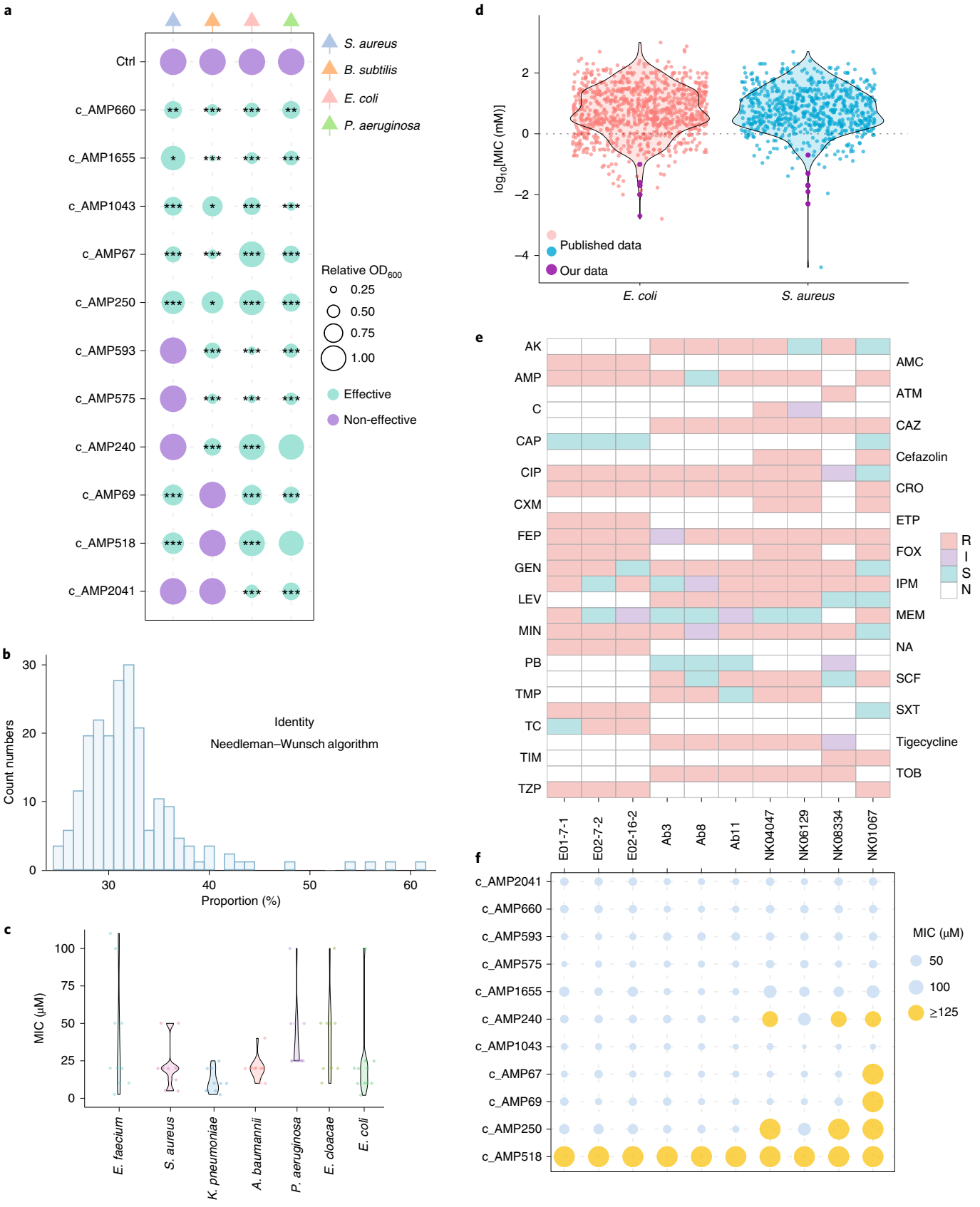
Although NLP methods currently take text (in our case, sequences) as input, future efforts could incorporate structural information that is critical for peptide functions. Currently, among our selected peptides, we observed an enrichment of helix-dominated peptides as revealed by circular dichroism (7/11; Extended Data Fig. 5), agreeing with previous findings of alpha-helix, pore-forming peptides among the most potent AMPs[21,54,55]. Another 11 randomly chosen peptides from the rest of the 181 peptides displayed highly diverse, complex structures. For the selected c_AMPs, their low MICs against *E. coli* and *S. aureus* are similar to the best-known AMPs, and the rich collection of molecules enables us to further select those with low cytotoxicity against eukaryotic cells for in vivo testing, eventually aiming to clinical use. Available MIC data in the curated AMP database revealed only seven AMPs with lower MIC toward *E. coli*; among those, no AMP was of bacterial origin, and only one animal cathelicidin from banded krait (*Bungarus fasciatus*) was tested to be effective on MDR, Gram-negative bacteria. Retrospectively, MIC values against a limited number of bacteria were available only for a small proportion of all AMPs and, thus, were not considered as a selection criterion in our study, leaving room for further pruning of AMP data for training or even mining a subgroup of highly potent AMPs if more data become available.

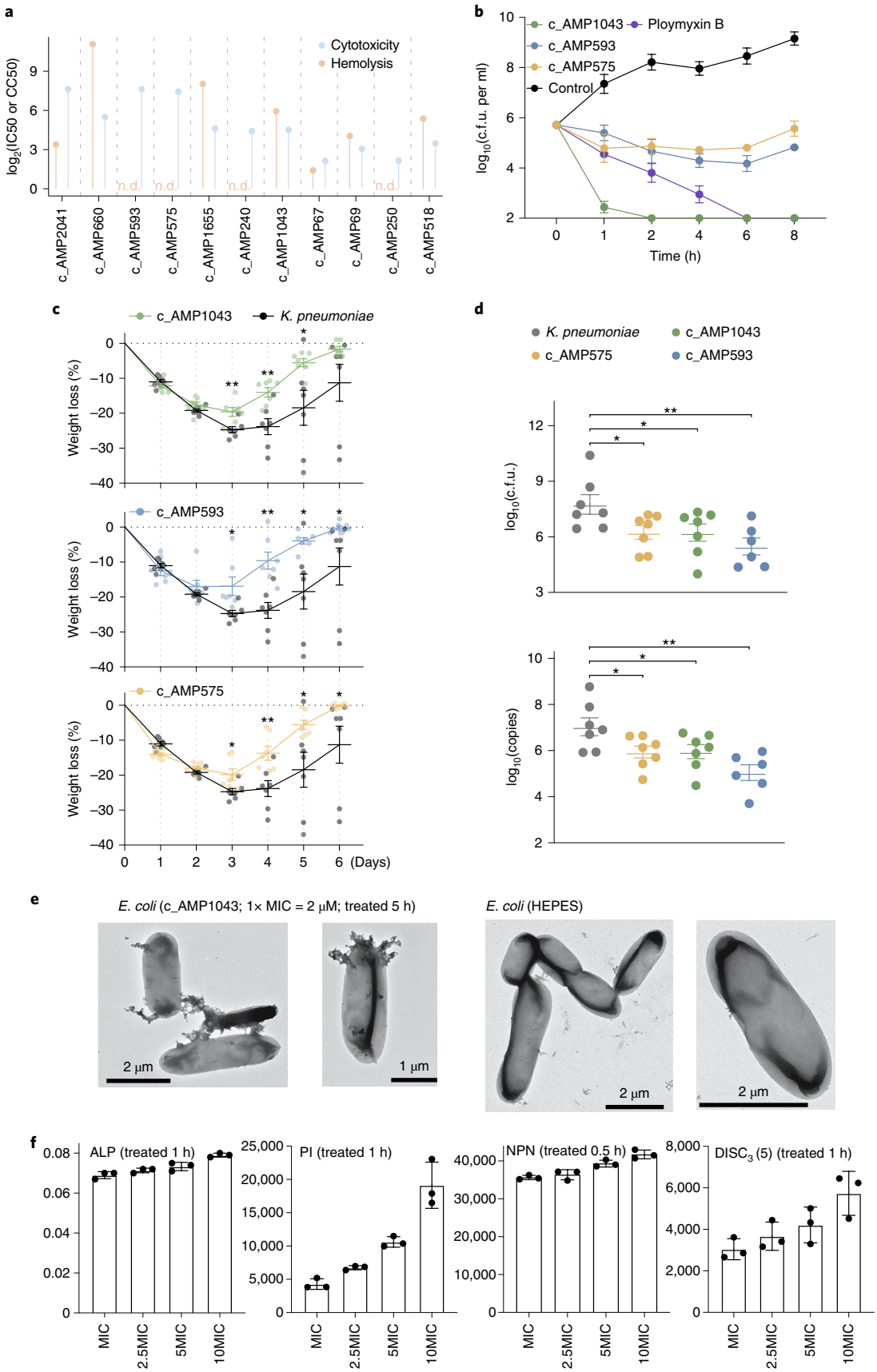
We eventually selected three peptides of low toxicity in assays using human cells and confirmed their effectiveness against typical lung bacterial infection by *K. pneumoniae* in a mouse model. These and other c_AMPs can serve as starting points for pharmaceutical optimization—for example, via peptide stapling[56] or chimerization[57]—to support further potency increases. Additionally, further in-depth profiling of the action mechanisms of these c_AMPs can also help increase their specificity, as, so far, all mechanisms discovered focused on the cell membrane. Moreover, such modifications could attenuate some of the toxicity against mammalian cells that precluded further validation of potent c_AMPs with in vivo assays; additionally, in broader industrial applications, this problem might be less critical for AMPs. In particular, c_AMP67, c_AMP69 and c_AMP660 showed activity against MDR, Gram-negative bacteria—one of the major infectious threats in clinics and with few effective antibiotics approved for

Fig. 4 | Experimental validation and potency assays of predicted AMPs. a, Spectra and level of bacterial inhibition of 11 short-listed c_AMPs against the four strains of bacteria used for the initial screening, showing that they are effective (green) against multiple species, and with high potency of bacterial inhibition (up to >75% reduction in OD_{600}, at a concentration of 60 μM). '*' means $0.01 < P \leq 0.05$; '**' means $0.001 < P \leq 0.01$; and '***' means $P \leq 0.001$ in Dunnett's test (two-sided). **b**, Distribution of highest similarity between 181 verified c_AMPs in our study to that of the training dataset. Most of our c_AMPs have less than 40% similarity to previously known AMPs in the training dataset. Data are presented as mean values ± s.e.m. **c**, Distribution of MICs of 11 selected c_AMPs against expanded list of pathogens; for *K. pneumoniae*, all c_AMPs have MIC < 25 μM; and for other Gram-negative species, including *A. baumanii, E. cloacae* and *E. coli*, at least one c_AMP reached MIC of 25 μM. At the same time, all c_AMPs reached <10 μM in MIC against one Gram-negative bacterial species. Data are presented as mean values ± s.e.m. **d**, The range of MICs obtained for our selected AMPs (purple) with reference to collected MIC values in the database, against *E. coli* (left, light red for known AMPs) and *S. aureus* (right, blue for known AMPs), demonstrating that our AMPs are among the most potent discovered so far. $n = 3$ independent experiments. Data are presented as mean values ± s.e.m. **e**, Antibiotic resistance profiles representative of MDR, Gram-negative (MDR G^-) bacteria strains, including *A. baumannii* (Ab3, Ab8 and Ab11), *K. pneumoniae* (NK04047, NK06129, NK08334 and NK01067) and *E. coli* (E01-7-1, E02-7-2 and E02-16-2). Antibiotic resistance: red means resistant (R), purple means susceptible increased exposure (I), green means susceptible (S) and blank block means not tested (N). Antibiotics that have been tested in those strains included Amikacin (AK), Amoxicillin-clavulanate (AMC), Ampicillin (AMP), Aztreonam (ATM), Chloramphenicol (C), Ceftazidime (CAZ), Chloramphenicol (CAP), Cefazolin, Ciprofloxacin (CIP), Cephalexin (CL), Ceftriaxone (CRO), Cefotaxime (CTX), Cefuroxime (CXM), Ertapenem (ETP), Cefepime (FEP), Cefoxitin (FOX), Gentamicin (GEN), Imipenem (IPM), Levofloxacin/Hemihydrate (LEV), Meropenem (MEM), Minocycline-HCL (MIN), Nalidixicacid (NA), Polymyxin-B (PB), Cefoperazone/sulbactam (SCF), Trimethoprim (TMP), Sulfamethoxazole (SXT), Tetracycline (TC), tigecycline, Ticarcillin (TIM), Tobramycin (TOB) and Piperacillin/tazobactam (TZP). **f**, Corresponding MICs of 11 selected c_AMPs against tested MDR G^- bacteria strains, demonstrating that multiple c_AMPs (such as c_AMP1043, c_AMP575, c_AMP595, c_AMP660 and c_AMP2043) are highly potent against all strains, and the rest of the c_AMPs, except c_AMP518, are potent against multiple strains. Data are presented as mean values ± s.e.m.

clinical use[58,59]. Thus, exploring modifications of the peptides that might reduce mammalian cytotoxicity while preserving antibacterial activity might merit further attention.

Overall, our study showcases the utility of combining NLP methods and large microbiome data for the mining of AMPs. Our proof-of-concept study yielded voluminous diverse AMPs





that confer their various antibacterial effects via divergent action mechanisms. Beyond providing this set of promising candidate molecules for further improvement for challenging pharmaceutical applications (for example, antibiotic-resistant bacteria), our study shows that an NLP-driven approach can achieve a high success rate for target class peptide identification in a much-shortened

Fig. 5 | c_AMP treatment of a mouse model of bacterial infection and mechanistic assays for c_AMP1043. a, Eukaryotic toxicity screen and hemolysis experiment determined CC50/IC50 for selected AMPs, shown as $\log_2$-transformed values. CC50 was determined using the MTT test on HCT116 cells with a range of concentrations (Methods); hemolysis was determined using selected AMPs and fresh human red blood cells and different concentrations of AMPs (Methods). ND, not determined—that is, no hemolytic effects. **b**, Time–kill assays of three c_AMPs used in the animal experiments. *E. coli* cells were then treated with 16× MIC AMPs in liquid media, and c.f.u. was measured at continual time points. Results indicate that the lowest c.f.u. values were achieved between 2 and 6 h. $n = 3$ in each group, with polymyxin B as positive control. **c**, Three c_AMPs without significant toxicity against mammalian cells were effective in treating a mouse model of *K. pneumoniae* infection. Mice were inoculated intra-nasally with 10^9 c.f.u. of *K. pneumoniae* (ATCC7 00603) strain first, and, on day 2, 50 µl (5× MIC) of c_AMPs were intra-nasally administered to infected mice, $n = 7$ for each group. Compared with the non-treated, *K. pneumoniae*-infected group, the c_AMP-treated group showed significantly faster recovery of body weight. $^*P < 0.05$ and $^{**}P < 0.01$ in two-sided *t*-test. **d**, Bacterial load significantly decreased in the lung tissues after being treated with c_AMPs. The c.f.u. was determined on standardized weight of tissues in *K. pneumoniae*-infected lungs and c_AMP-treated lungs, using Mueller–Hinton agar (CaMHA) (Methods). $n = 6$–7 per group; '*' denotes $P < 0.05$ and '**' denotes $P < 0.01$ in two-sided *t*-test. **e**, TEM examination of *E. coli* cells treated with c_AMP1043 and HEPES, showing cell content leakage and disruption of cell wall/membrane; experiments were performed in triplicate with similar results, and one representative figure is shown. **f**, Mechanistic tests, including ALP, PI, NPN and $DISC_3(5)$ assays, revealed that c_AMP1043 is capable of functioning via disruption of the outer membrane of Gram-negative bacteria *E. coli*. In PI, NPN and DISC3(5) assays, dosage-dependent increase of signals was observed, reflecting the effects of c_AMP1043 on corresponding cell components. $n = 3$ for each assay.

timeframe compared to traditional experiment-based methods. This approach also represents a 're-purposing' application for the many sequencing-based studies of environmental and medical metagenomics with large-scale datasets, to discover a part of the functional 'dark matter'. It is clear that the application of this approach can greatly facilitate the identification and prioritization of peptide agents for research and for therapies. Lastly, it bears emphasizing that very similar approaches should be suitable for mining other types of peptides involved in microbial signaling and for modulating host immunity or metabolism.

Online content

Any methods, additional references, Nature Research 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/s41587-022-01226-0.

Received: 6 July 2021; Accepted: 19 January 2022;
Published online: 3 March 2022

References

1. O'Neil, J. *Tackling drug-resistant infections globally: final report and recommendations.* (Review on Antimicrobial Resistance, 2016).
2. De Oliveira, D. M. P. et al. Antimicrobial resistance in ESKAPE pathogens. *Clin. Microbiol. Rev.* **33**, e00102-19 (2020).
3. Tacconelli, E. et al. *Global priority list of antibiotic-resistant bacteria to guide research, discovery, and development of new antibiotics.* (World Health Organization, 2017).
4. PEW Charitable Trusts. Analysis shows continued deficiencies in antibiotic developments since 2014. *PEW* https://www.pewtrusts.org/en/research-and-analysis/data-visualizations/2019/five-year-analysis-shows-continued-deficiencies-in-antibiotic-development (2019).
5. Lazzaro, B. P., Zasloff, M. & Rolff, J. Antimicrobial peptides: application informed by evolution. *Science* **368**, eaau5480 (2020).
6. Heng, N. C. K. & Tagg, J. R. What's in a name? Class distinction for bacteriocins. *Nat. Rev. Microbiol.* **4**, 160–160 (2006).
7. Chen, X. et al. Roles and mechanisms of human cathelicidin LL-37 in cancer. *Cell. Physiol. Biochem.* **47**, 1060–1073 (2018).
8. Yu, G., Baeder, D. Y., Regoes, R. R. & Rolff, J. Predicting drug resistance evolution: insights from antimicrobial peptides and antibiotics. *Proc. Biol. Sci.* **285**, 20172687 (2018).
9. Kintses, B. et al. Phylogenetic barriers to horizontal transfer of antimicrobial peptide resistance genes in the human gut microbiota. *Nat. Microbiol.* **4**, 447–458 (2019).
10. Buffie, C. G. & Pamer, E. G. Microbiota-mediated colonization resistance against intestinal pathogens. *Nat. Rev. Immunol.* **13**, 790–801 (2013).
11. Bisanz, J. E. et al. A genomic toolkit for the mechanistic dissection of intractable human gut bacteria. *Cell Host Microbe* **27**, 1001–1013 (2020).
12. Wilson, M. R. et al. The human gut bacterial genotoxin colibactin alkylates DNA. *Science* **363**, eaar7785 (2019).
13. Kent, A. G., Vill, A. C., Shi, Q., Satlin, M. J. & Brito, I. L. Widespread transfer of mobile antibiotic resistance genes within individual gut microbiomes revealed through bacterial Hi-C. *Nat. Commun.* **11**, 4379 (2020).
14. Sberro, H. et al. Large-scale analyses of human microbiomes reveal thousands of small, novel genes. *Cell* **178**, 1245–1259 (2019).
15. Kim, S. G. et al. Microbiota-derived lantibiotic restores resistance against vancomycin-resistant *Enterococcus*. *Nature* **572**, 665–669 (2019).
16. Li., J. et al. Mining the human tonsillar microbiota as autoimmune modulator. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/719807v1.full (2019).
17. Walsh, C. T. Insights into the chemical logic and enzymatic machinery of NRPS assembly lines. *Nat. Prod. Rep.* **33**, 127–135 (2016).
18. Spänig, S. & Heider, D. Encodings and models for antimicrobial peptide classification for multi-resistant pathogens. *BioData Min.* **12**, 7 (2019).
19. Stokes, J. M. et al. A deep learning approach to antibiotic discovery. *Cell* **180**, 688–702 (2020).
20. Zhavoronkov, A. et al. Deep learning enables rapid identification of potent DDR1 kinase inhibitors. *Nat. Biotechnol.* **37**, 1038–1040 (2019).
21. Das, P. et al. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nat. Biomed. Eng.* **5**, 613–623 (2021).
22. Nagarajan, D. et al. Computational antimicrobial peptide design and evaluation against multidrug-resistant clinical isolates of bacteria. *J. Biol. Chem.* **293**, 3492–3509 (2018).
23. Van Oort, C. M., Ferrell, J. B., Remington, J. M., Wshah, S. & Li, J. AMPGAN v2: machine learning-guided design of antimicrobial peptides. *J. Chem. Inf. Model.* **61**, 2198–2207 (2021).
24. Wang, C., Garlick, S. & Zloh, M. Deep learning for novel antimicrobial peptide design. *Biomolecules* **11**, 471 (2021).
25. Gupta, A. & Zou, J. Feedback GAN for DNA optimizes protein functions. *Nat. Mach. Intell.* **1**, 105–111 (2019).
26. Veltri, D., Kamath, U. & Shehu, A. Deep learning improves antimicrobial peptide recognition. *Bioinformatics* **34**, 2740–2747 (2018).
27. Devlin., J., Chang., M.-W., Lee., K. & Toutanova., K. BERT: pre-training of deep bidirectional transformers for language understanding. Preprint at https://arxiv.org/abs/1810.04805 (2018).
28. Pasolli, E. et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell* **176**, 649–662 (2019).
29. Bevins, C. L. & Salzman, N. H. Paneth cells, antimicrobial peptides and maintenance of intestinal homeostasis. *Nat. Rev. Microbiol.* **9**, 356–368 (2011).
30. Login, F. H. et al. Antimicrobial peptides keep insect endosymbionts under control. *Science* **334**, 362–365 (2011).
31. World Health Organization. *2019 Antibacterial Agents in Clinical Development* (World Health Organization, 2019).
32. Gong, L. et al. A nosocomial respiratory infection outbreak of carbapenem-resistant *Escherichia coli* ST131 with multiple transmissible bla_{KPC-2} carrying plasmids. *Front. Microbiol.* **11**, 2068 (2020).
33. Upert, G., Luther, A., Obrecht, D. & Ermert, P. Emerging peptide antibiotics with therapeutic potential. *Med. Drug Discov.* **9**, 100078 (2021).
34. Cigana, C. et al. Efficacy of the novel antibiotic POL7001 in preclinical models of *Pseudomonas aeruginosa* pneumonia. *Antimicrob. Agents Chemother.* **60**, 4991–5000 (2016).
35. Florin, T. et al. An antimicrobial peptide that inhibits translation by trapping release factors on the ribosome. *Nat. Struct. Mol. Biol.* **24**, 752–757 (2017).

36. Gagnon, M. G. et al. Structures of proline-rich peptides bound to the ribosome reveal a common mechanism of protein synthesis inhibition. *Nucleic Acids Res.* **44**, 2439–2450 (2016).
37. Chu, H. et al. Human α-defensin 6 promotes mucosal innate immunity through self-assembled peptide nanonets. *Science* **337**, 477–481 (2012).
38. Loth, K. et al. The ancestral N-terminal domain of big defensins drives bacterially triggered assembly into antimicrobial nanonets. *mBio* **10**, e01821-19 (2019).
39. De Fauw, J. et al. Clinically applicable deep learning for diagnosis and referral in retinal disease. *Nat. Med.* **24**, 1342–1350 (2018).
40. Coudray, N. et al. Classification and mutation prediction from non-small cell lung cancer histopathology images using deep learning. *Nat. Med.* **24**, 1559–1567 (2018).
41. Kather, J. N. et al. Deep learning can predict microsatellite instability directly from histology in gastrointestinal cancer. *Nat. Med.* **25**, 1054–1056 (2019).
42. Xiong, Z. et al. Pushing the boundaries of molecular representation for drug discovery with the graph attention mechanism. *J. Med. Chem.* **63**, 8749–8760 (2020).
43. Zhong, H. et al. Distinct gut metagenomics and metaproteomics signatures in prediabetics and treatment-naïve type 2 diabetics. *EBioMedicine* **47**, 373–383 (2019).
44. Fjell, C. D., Hancock, R. E. & Cherkasov, A. AMPer: a database and an automated discovery tool for antimicrobial peptides. *Bioinformatics* **23**, 1148–1155 (2007).
45. Zhao, X., Wu, H., Lu, H., Li, G. & Huang, Q. LAMP: a database linking antimicrobial peptides. *PLoS ONE* **8**, e66557 (2013).
46. Chu, J., Vila-Farres, X. & Brady, S. F. Bioactive synthetic-bioinformatic natural product cyclic peptides inspired by nonribosomal peptide synthetase gene clusters from the human microbiome. *J. Am. Chem. Soc.* **141**, 15737–15741 (2019).
47. Garcia-Gutierrez, E., Mayer, M. J., Cotter, P. D. & Narbad, A. Gut microbiota as a source of novel antimicrobials. *Gut Microbes* **10**, 1–21 (2019).
48. Ryu, M., Park, J., Yeom, J. H., Joo, M. & Lee, K. Rediscovery of antimicrobial peptides as therapeutic agents. *J. Microbiol.* **59**, 113–123 (2021).
49. Cullen, T. W. et al. Gut microbiota. Antimicrobial peptide resistance mediates resilience of prominent gut commensals during inflammation. *Science* **347**, 170–175 (2015).
50. Müller, A. T., Gabernet, G., Hiss, J. A. & Schneider, G. modlAMP: Python for antimicrobial peptides. *Bioinformatics* **33**, 2753–2755 (2017).
51. Agrawal, P. & Raghava, G. P. S. Prediction of antimicrobial potential of a chemically modified peptide from its tertiary structure. *Front. Microbiol.* **9**, 2551 (2018).
52. Lertampaiporn, S., Vorapreeda, T., Hongsthong, A. & Thammarongtham, C. Ensemble-AMPPred: robust AMP prediction and recognition using the ensemble learning method with a new hybrid feature for differentiating AMPs. *Genes* **12**, 137 (2021).
53. Barrett, R., Jiang, S. & White, A. D. Classifying antimicrobial and multifunctional peptides with Bayesian network models. *Pept. Sci.* **110**, e24079 (2018).
54. Kumar, P., Kizhakkedathu, J. N. & Straus, S. K. Antimicrobial peptides: diversity, mechanism of action and strategies to improve the activity and biocompatibility in vivo. *Biomolecules* **8**, 4 (2018).
55. Guha, S., Ghimire, J., Wu, E. & Wimley, W. C. Mechanistic landscape of membrane-permeabilizing peptides. *Chem. Rev.* **119**, 6040–6085 (2019).
56. Mourtada, R. et al. Design of stapled antimicrobial peptides that are stable, nontoxic and kill antibiotic-resistant bacteria in mice. *Nat. Biotechnol.* **37**, 1186–1197 (2019).
57. Luther, A. et al. Chimeric peptidomimetic antibiotics against Gram-negative bacteria. *Nature* **576**, 452–458 (2019).
58. Munoz-Price, L. S. et al. Clinical epidemiology of the global expansion of *Klebsiella pneumoniae* carbapenemases. *Lancet Infect. Dis.* **13**, 785–796 (2013).
59. Bonomo, R. A. et al. Carbapenemase-producing organisms: a global scourge. *Clin. Infect. Dis.* **66**, 1290–1297 (2018).
60. Santos-Júnior, C. D., Pan, S., Zhao, X. M. & Coelho, L. P. Macrel: antimicrobial peptide screening in genomes and metagenomes. *PeerJ* **8**, e10555 (2020).
61. Bhadra, P., Yan, J., Li, J., Fong, S. & Siu, S. W. I. AmPEP: sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. *Sci. Rep.* **8**, 1697 (2018).
62. Xiao, X., Wang, P., Lin, W. Z., Jia, J. H. & Chou, K. C. iAMP-2L: a two-level multi-label classifier for identifying antimicrobial peptides and their functional types. *Anal. Biochem.* **436**, 168–177 (2013).
63. Meher, P. K., Sahu, T. K., Saini, V. & Rao, A. R. Predicting antimicrobial peptides with improved accuracy by incorporating the compositional, physico-chemical and structural features into Chou's general PseAAC. *Sci. Rep.* **7**, 42362 (2017).
64. Fingerhut, L., Miller, D. J., Strugnell, J. M., Daly, N. L. & Cooke, I. R. ampir: an R package for fast genome-wide prediction of antimicrobial peptides. *Bioinformatics* **36**, 5262–5263 (2020).
65. Xiao, X., Shao, Y. T., Cheng, X. & Stamatovic, B. iAMP-CA2L: a new CNN-BiLSTM-SVM classifier based on cellular automata image for identifying antimicrobial peptides and their functional types. *Brief. Bioinform.* **22**, bbab209 (2021).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature America, Inc. 2022

Methods

Data collection. We collected, in total, five kinds of datasets in our study, including AMPs, non-AMPs, representative bacterial genomes, metaproteome and large-scale metagenome.

AMPs. AMP data were mainly collected from four public AMP datasets—ADAM[26], APD[66], CAMP[67] and LAMP[45]—which cover most of AMP sequences from different sources (downloaded as of 2 October 2018). These four datasets were merged into a large dataset, and the duplicated sequence in the dataset was removed, resulting in 10,327 AMP sequences, and the length of these sequences range from 6 AAs to 518 AAs. Five sequences whose lengths are longer than 300 AAs were removed from the dataset, resulting in 10,322 non-redundant AMPs, 55.17% of which were shorter than 50 AAs (Fig. 2a). Note that peptides with only anticancer activity are not our target (AMPs contain antibacterial, antiviral and antifungal activity[5]), and, in the database, only 0.15% (15 of 10,327 AMPs) were labeled with anticancer; however, all of these AMPs also have antibacterial activity. Eventually, we got 10,322 AMP sequences in total for subsequent analysis.

Non-AMPs. The non-AMP dataset was downloaded from UniProt (http://www.uniprot.org) by setting the 'subcellular location' filter to cytoplasm and removing any entry that matched the following keywords: antimicrobial, antibiotic, antiviral, antifungal, effector or excreted (downloaded as of 20 November 2018). Then, we deleted the duplicated sequences in the dataset and kept only peptides shorter than 300 AAs. Meanwhile, the sequences identical to any AMPs were removed from the non-AMP dataset so there were 3,029,894 non-AMP sequences left, among which 114,995 were shorter than 50 AAs (Fig. 2a).

Training dataset. The training dataset was used for building NLP models. The AMP dataset was first split into two sets at a ratio of 8:2, with 8,290 AMPs kept in the training dataset. The non-AMP training dataset contained 74,838 sequences (with matched distribution to that of the AMP training dataset; Extended Data Fig. 1a).

Test dataset. The test dataset was used to evaluates the performance of NLP models. The AMPs in the test dataset and the training set are independent of each other, as are the non-AMPs; a total of 2,032 and 2,908,751 sequences were kept in AMP and non-AMP, respectively. The peptides in the test dataset with lengths of 50 AAs or fewer were taken as the final test dataset, which contained 1,085 AMPs and 58,776 non-AMPs.

Validation datasets. We collected 677 bacteriocins from the BAGEL4 database[68] as AMPs and a total of 5,541 non-AMPs from the Neme et al. study[69] (downloaded as of 30 September 2018). This part of the data is not included in the training or test set and used only to tune hyperparameters.

Representative genomes. The representative genomes dataset was derived from species-level genome bins, which contain 154,723 genomes[28], and was used to predicted sORF sequences. We kept only the bins whose completeness was greater than 90%, which filters out more than half of the bins (Fig. 3a). We constructed a species-level resolution, non-redundant representative genomes dataset. To achieve this, we defined a selection score (= 0.7 × completeness − 0.1 × number of contigs − 0.3 × contamination) and selected the one with the highest score when multiple bins were present for the same taxonomic classification. For bins that cannot be assigned to the species level, we used kr (version 2.0.2)[70] to calculate pairwise distances between bins, and those with distance less than 0.1 were merged into the same species-level operational taxonomic units, with the genomic bin with the highest selection score as representative. The final representative genomes dataset contains 4,409 high-quality species-level annotations.

Metaproteome. In this study, to ensure that our sORFs are indeed expressed, multiple metaproteome datasets were used to select AMP sequences. These metaproteome datasets were collected from Gavin et al.[71], Tanca et al.[72], Sandip et al.[73], Young et al.[74] and Zhong et al.[43]. We combined these five metaproteome datasets and selected the sequences with lengths shorter than 50 AAs, resulting in 19,881,257 peptides.

Large-scale metagenome. To identify the AMPs that have the inhibitory potential toward either a wide range of bacteria or specific groups, we established correlations between AMPs and bacterial taxa in 15 independent, large-scale metagenomic cohorts[75–89]. Among those, 11 cohorts contain at least 100 human gut microbiome samples, and all cohorts contain, in total, 11,011 samples.

Establishment and curation of NNMs. We trained, in total, five NNMs to distinguish AMPs from non-AMPs. To start with, we converted the AMP and non-AMP datasets into a fixed-size vector, and the 20 basic AAs were converted to 1~20 numerical form (details in Supplementary Table 16). Sequence vectors are padded with 0 if the raw sequence did not reach 300 AAs. The sequence vectors add the number 1/0 in the last column as the classification label of the sequences, indicating AMPs/non-AMPs, respectively.

The most fundamental model is a convolutional NNM with LSTM layer architecture, which has been used in the current research of Veltri et al.[26]. Details of this architecture were as follows: the embedding layer (input_dim: 21, output_dim: 128 and input_length: 300); the 1D Conv layer (nb_filter: 64, filter_length: 16, strides: 1, activation: relu); 1D max pooling layer (pool_size: 5, strides: 5); LSTM layer (units: 100, unroll: True, stateful: False); and dense layer (units: 1, activation: sigmoid). The second model was created by changing the LSTM layer to the Attention layer in the same NNM architecture of the first model[90]. The Attention layer can effectively capture long-range dependencies[91]—connections between any two (or more) AAs in the whole protein sequences, instead of only neighboring AAs. This model has no adding extra input information; further parameters for other layers were units = 100, unroll = True, stateful = False; epochs = 200, batch size = 20, learning rate = 0.001. ADAM optimizer (with beta_1 = 0.9, beta_2 = 0.999, epsilon = None, decay = 0.0, amsgrad = False) was used for ATT/LSTM model optimization, with loss function Cross entropy; dropout was set to 0.1 after MaxPooling1D layer. Both models used the same amount of non-AMPs (8,290), subsisted from 74,838 non-AMPs, to first establish balanced training models as baselines (b_LSTM and b_ATT). Then, we used the higher amount of non-AMPs (all 74,838) in the training data to train the two models mentioned above to obtain our third and fourth classification models (LSTM and ATT). All four models were built with the Keras framework (version 2.2.4, https://www.keras.io) using a sequential model and a TensorFlow (version 1.14.0)[92] deep learning library back-end. The last model was a pre-trained representation model applied in NLP called BERT. BERT learns contextual information from an unsupervised corpus and generates corresponding representation vectors; it has been recognized for its wide adaptability to different NLP downstream tasks, including text classification and sequence annotation. This model was fine-tuned from a pre-trained model (bert-base-uncased) provided in the original study as weights that have been initialized well. BERT was trained with PyTorch (version 1.0.1)[93]. We treated AAs as text information with each AA as a one-word code. During training, AAs were separated by gaps, and sequence beginnings/ends were marked with [CLS] and [SEP] labels. We then added a linear layer at the end of the BERT model to reduce the dimension to 2. Initial parameters from 'bert-base-uncased' were then fine-tuned with Cross entropy as loss function, ADAM with default parameters as optimizer, batch size = 64 and learning rate = 2×10^{-5}. To prevent over-fitting, we trained with 'early-stop' strategy, in that models were stopped and saved when they started to decrease in performance. Additional ten-fold cross-validation was applied. All models converged quickly during training. For individual models, they were fixed and unchanged after the training, and a peptide with prediction score greater than 0.5 (positive) is considered a candidate AMP.

For hyperparameters, we used a third, albeit small, dataset (667 AMPs and 5,541 non-AMPs) for their choices, namely batch_size and learning_rate. For BERT, we tuned the learning_rate between 1×10^{-6} and 5×10^{-5} and the batch_size between 32 and 128. For Attention and LSTM models, we used dynamic learning_rate: new_lr = $0.9 \times \mathrm{lr}^{\frac{\mathrm{steps}}{5}}$, 0.9 being the decay rate, 5 being the decay steps, steps = global_step, lr = learning_rate starting at 0.001; batch_size was tested at 10, 20 and 40. This has set the optimal batch_size for all models and optimal learning_rate for BERT.

Prediction pipeline by combination of multiple models. We used the test dataset that is composed of 2,908,751 non-AMPs and 2,032 AMPs as input data of these five classification models and made the results output as vectors. These result vectors include before-training information about a series of prediction features of these five classification models, in particular hidden layers of the NNMs as well as their processed features of input sequences. We extracted the last hidden layer (except fully connected layer) of values of NNMs from Attention, LSTM or BERT, reflecting their individual properties in sequence classifications. For each sequence, the output vector of size 1 × N was obtained, without any processed information added. N in each model is determined during building of the models: for BERT, N = 768; for Attention models, N = 64; and for LSTM, N = 100. Each model reached different numbers of TPs and FPs: outside of 2,032 AMPs and 2,908,751 non-AMPs in test sets, BERT has TP = 1,777 and FP = 19,408; LSTM has TP = 1,968 and FP = 624,417; b_LSTM has TP = 1,917 and FP = 98,339; ATT has TP = 1,827 and FP = 20,185; and b_ATT has TP = 1,909 and FP = 107,375. We chose the TPs and FPs shared by all models (TP: 1,678 and FP: 3,981) and calculated pairwise correlations (ten pairs) of the same sequence among the five models.

We defined evaluation parameters to evaluate the performance of different model combinations. The first, Precision of intersection, is defined as follows:

$$\bigcap_{i=1}^{n} \mathrm{Precision}_i = \frac{\bigcap_{i=1}^{n} \mathrm{TP}_i}{\bigcap_{i=1}^{n} \mathrm{TP}_i + \bigcap_{i=1}^{n} \mathrm{FP}_i}$$

Here, the letter n in the equation represents the number of models; the symbol ∩ means the intersection algorithm of models; and TP and FP stand for true positive and false positive, respectively. The common Precision is defined as TP/(TP + FP); however, the TP and FP were exchanged by ∩TP and ∩FP in our Precision.

$$\mathrm{Recall} = \frac{\bigcap_{i=1}^{n} \mathrm{TP}_i}{\mathrm{CP}}$$

The second, Recall, is defined as follows:

CP refers to condition positive. Our new parameter resembles the parameter Recall, which has also been called sensitivity. Commonly, Recall = TP/CP, and we used intersection TP instead of TP.

We used this dataset as a benchmark for the different AMP prediction methods, using AUPRCs as evaluation parameters for these models in addition to the Precision, using the function of precision_recall_curve from the sklearn.metrics module (Python 3). The relevant scripts are available on GitHub.

Prediction of sORFs and candidate AMPs. We used the 'getorf' function of the EMBOSS software package (version 6.6.0.0) to predict sORFs from representative genomes. The parameters were set as '-find 2 -table 11 -minsize 15 -maxsize 150'[94]. Then, the sORF dataset was predicted by our pipeline to identify c_AMPs.

Selection of protein potential AMPs with multi-omics. To ensure that sORFs predicted in silico are likely expressed to proteins/peptides, we cross-examined with metaproteomic datasets. We calculated the *k*-mers of all sORF sequences (where *k* is at least half the length of the sequence, and the maximum value is the original length of the sequence) and checked for peptide sequences in metaproteomics data. If there is a perfect match between a *k*-mer and a metaproteomic peptide, it indicated that more than half of the sequence of a specific sORF is present as a peptide in proteomics data, which provides additional evidence that this sORF is likely to be expressed.

c_AMP profiling in gut metagenomic samples. To obtain abundance information of c_AMPs in each metagenomic sample, we used PALADIN software (version 1.4.0) to align c_AMP sequences with metagenomics reads[95] and then used SAMtools (version 1.7) to calculate the abundance of c_AMPs (functions of SAMtools used included 'sort', 'index' and 'idxstats')[96] by calculating their coverage per million reads.

Taxonomic profiling of gut metagenomic samples. Relative abundances of bacteria corresponding to representative genomes were calculated by MetaPhlAn2 (ref. [97]). The unique clade-specific marker genes used as reference in MetaPhlAn2 were updated by our selected representative genome, and all other parameters remained unchanged.

Network analysis and candidate AMP selection. The correlation between taxonomical composition (genus and species level for each sample) and c_AMPs was calculated by the corAndPvalue (Spearman) function from the R package WGCNA (version 1.68)[98,99]. We selected c_AMPs (ranging from 74 to 1,599; Supplementary Table 17) and bacterial genera/species that appear in more than 5% of the samples in each cohort, and correlational *P* value was adjusted to FDR using the function 'mt.rawp2adjp' (Benjamini–Hochberg correction) of the multtest package in R[100] (version 3.4.1; FDR threshold: 0.05). The screened correlation pairs were kept when they appeared in at least seven cohorts, and correlational network was visualized using Cytoscape (version 3.6.1)[101].

Sequence similarity estimation. We applied the Needleman–Wunsch algorithm in the function 'needleall' from the EMBOSS software package (version 6.6.0.0) to estimate the similarity between our c_AMPs and AMPs in the training dataset—first, by alignment, and second, by counting the identical AA pairs in the alignment. The parameters used are all default, and the parameter 'identity' was sifted out for the graph.

AA frequencies. The function 'ProteinAnalysis' was used to calculate the AA content in percentages of each peptide. This function was imported from the Biopython module 'Bio.SeqUtils.ProtParam' (version 1.75)[102].

Peptide synthesis. The peptides used in this study were synthesized by solid-phase peptide synthesis[102] by Royo Biotech, and their accurate molecular weights were determined by mass spectrometry. The purity of all peptides was determined by high-performance liquid chromatography, and all purity was greater than 90%.

Bacterial inhibition experiment. Four strains of bacteria—including representative Gram-positive bacteria *B. subtili* (ATCC 23857) and *S. aureus* (ATCC 6538) and representative Gram-negative bacteria *E. coil* DH5α and *P. aeruginosa* (ATCC 15692)—were streaked on Luriae–Bertani (LB) agar medium and incubated at 37 °C overnight. The individual colonies were picked into LB culture medium and shaken at 120 r.p.m. at 37 °C overnight. The LB bacterial suspension was diluted to the predetermined starting concentration (optical density at 600 nm (OD_{600}) = 0.1) and then again diluted 1,000 times for the inhibition test.

We thawed freeze-dried powder of c_AMPs and dissolved in double-distilled water to 2.4 mmol L^{-1}. We set three experimental groups to test c_AMP antibacterial activity: (1) blank control group, 200 µl of LB solution; (2) bacterial control group, 100 µl of LB solution and 100 µl of bacterial solution; and (3) low-c_AMPs group (60 µmol L^{-1}), 100 µl of LB solution, 95 µl of bacterial solution and 5 µl of c_AMP solution. Experiments were performed on 96-well plates with each single well containing 200 µl of final volume. The OD_{600} value of each well was measured after culture at 37 °C for 12 h, and we normalized OD as calculated by (experimental group OD – blank OD) / (control group OD – blank OD). We used Prism software (version 8.4.0) to perform Dunnett's test for comparing experimental groups with a control group (two-sided). All experiments were performed with three independent replicates.

MIC determination. MIC determination of AMPs was performed by broth microdilution as described in Clinical and Laboratory Standards Institute guidelines[103]. In brief, *E. coli* strains (DH5α, E01-7-1, E02-7-2 and E02-16-2), *P. aeruginosa* (ATCC 15692,) *A. baumannii* strains (ATCC 19606, Ab3, Ab8 and Ab11), *K. pneumoniae* strains (NCTC 5056, NK67, NK04047, NK06129 and NK08334), *B. subtilis* (ATCC 23857), *S. aureus* (ATCC 6538) and *E. cloacae* (ATCC 13047) were inoculated in cation-adjusted Mueller–Hinton broth (CaMHB, QDRS Biotec, cat. no. 11865) at 37 °C overnight. The cultures were diluted 1:100 using the fresh CaMHB and subsequently cultured to the exponential phase (OD_{600} of 0.4–0.6), and then the cell concentration was adjusted to approximately 5×10^5 c.f.u. per milliliter, and 100-µl aliquots were transferred into 96-well plates containing 100 µl of different AMP solutions diluted serially two-fold. The tested ranges of AMPs were from 500 µM to 0.98 µM. After incubating at 37 °C for 16–18 h, the MIC was determined as the minimum concentration of c_AMPs where bacteria showed no detectable growth. For *E. faecium* (ATCC 19434), the same assay was conducted under anaerobic conditions using Brain-Heart Infusion Broth as the test medium. All assays were performed in triplicate. All experiments were performed with three independent replicates.

Cytotoxicity against mammalian cells. Cytotoxicity of c_AMPs was determined using the MTT Cell Proliferation and Cytotoxicity Assay Kit (Solarbio). Exponentially growing HCT116 cells were seeded in 96-well microtiter plates in appropriate cell culture medium. After 24-h incubation at 37 °C with 5% CO_2 in atmosphere, medium was replaced with fresh medium, and c_AMPs of various concentrations were added, followed by 48-h incubation. Each c_AMP experiment was performed with four independent replicates. Cell viability was monitored by the addition of MTT solution and measurement of optical density at OD_{490} after 4 h. Each c_AMP experiment was performed with three replicates. The CC50 value of each c_AMP was calculated by using online tools: https://www.aatbio.com/tools/ic50-calculator.

Hemolysis effect of c_AMPs. Fresh collected human red blood cells were first washed with PBS until the upper phase was clear after centrifugation (2,000 r.p.m.) and allocated onto 96-well U-bottom plates. Each c_AMP was diluted and added to the wells of various concentrations. After 1 h at 37 °C, cells were centrifuged at 3,000 r.p.m. for 10 min. The supernatant was diluted, and OD_{570} was measured. All experiments were performed with three independent replicates. The IC50 value of each c_AMP was calculated using online tools: https://www.aatbio.com/tools/ic50-calculator.

Additional method details on time-dependent killing, resistance, mouse model and biochemical assays, including PI, $DiSC_3(5)$, NPN and ALP, TEM and circular dichroism, can be found in the Supplementary Information.

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

Data availability

Our study contains only publicly available AMP, non-AMP, metagenome and metaproteome data. AMP data were mainly collected from four public AMP datasets—ADAM: http://bioinformatics.cs.ntou.edu.tw/adam/, APD: http://aps.unmc.edu, CAMP: http://www.camp.bicnirrh.res.in/ and LAMP: http://biotechlab.fudan.edu.cn/database/lamp/—which cover most of AMP sequences from different sources (downloaded as of 2 October 2018). The non-AMP dataset was downloaded from UniProt (https://www.uniprot.org) by setting the 'subcellular location' filter to cytoplasm and removing any entry that matches the following keywords: antimicrobial, antibiotic, antiviral, antifungal, effector or excreted (downloaded as of 20 November 2018). Validation datasets: non-AMPs part ENA project ID is PRJEB19640; AMPs part was downloaded from http://bagel4.molgenrug.nl/index.php. The representative genomes dataset was derived from species-level genome bins: https://opendata.lifebit.ai/table/SGB. The metaproteome datasets were collected from https://www.ebi.ac.uk/pride, PRIDE project IDs: PXD005780, PXD008870, PXD003907 and PXD000114. The 15 independent, large-scale metagenomic cohorts—BioProject IDs: PRJNA422434, PRJEB4336, PRJEB1220, PRJEB6337, PRJEB6456, PRJEB10878, PRJEB11532, PRJNA319574, PRJEB9584, PRJNA290380, PRJEB6337, PRJEB15371, PRJNA356102 and https://github.com/MetaSUB/MetaSUB-metadata. Source data are provided with this paper.

Code availability

The c_AMP prediction codes can be found at https://github.com/mayuefine/c_AMPs-prediction.

References

66. Wang, G., Li, X. & Wang, Z. APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res.* **44**, D1087–D1093 (2016).
67. Waghu, F. H., Barai, R. S., Gurung, P. & Idicula-Thomas, S. CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides. *Nucleic Acids Res.* **44**, D1094–D1097 (2016).
68. van Heel, A. J., de Jong, A., Montalbán-López, M., Kok, J. & Kuipers, O. P. BAGEL3: automated identification of genes encoding bacteriocins and (non-)bactericidal posttranslationally modified peptides. *Nucleic Acids Res.* **41**, W448–W453 (2013).
69. Neme, R., Amador, C., Yildirim, B., McConnell, E. & Tautz, D. Random sequences are an abundant source of bioactive RNAs or peptides. *Nat. Ecol. Evol.* **1**, 0217 (2017).
70. Domazet-Loso, M. & Haubold, B. Efficient estimation of pairwise distances between genomes. *Bioinformatics* **25**, 3221–3227 (2009).
71. Gavin, P. G. et al. Intestinal metaproteomics reveals host-microbiota interactions in subjects at risk for type 1 diabetes. *Diabetes Care* **41**, 2178–2186 (2018).
72. Tanca, A., Palomba, A., Pisanu, S., Addis, M. F. & Uzzau, S. Enrichment or depletion? The impact of stool pretreatment on metaproteomic characterization of the human gut microbiota. *Proteomics* **15**, 3474–3485 (2015).
73. Chatterjee, S. et al. A comprehensive and scalable database search system for metaproteomics. *BMC Genomics* **17**, 642 (2016).
74. Young, J. C. et al. Metaproteomics reveals functional shifts in microbial and human proteins during a preterm infant gut colonization case. *Proteomics* **15**, 3463–3473 (2015).
75. Danko, D. et al. A global metagenomic map of urban microbiomes and antimicrobial resistance. *Cell* **184**, 3376–3393 (2021).
76. Zeevi, D. et al. Personalized nutrition by prediction of glycemic responses. *Cell* **163**, 1079–1094 (2015).
77. Vatanen, T. et al. Variation in microbiome LPS immunogenicity contributes to autoimmunity in humans. *Cell* **165**, 842–853 (2016).
78. Schirmer, M. et al. Linking the human gut microbiome to inflammatory cytokine production capacity. *Cell* **167**, 1125–1136 (2016).
79. Bäckhed, F. et al. Dynamics and stabilization of the human gut microbiome during the first year of Life. *Cell Host Microbe* **17**, 690–703 (2015).
80. Nielsen, H. B. et al. Identification and assembly of genomes and genetic elements in complex metagenomic samples without using reference genomes. *Nat. Biotechnol.* **32**, 822–828 (2014).
81. Qin, J. et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. *Nature* **490**, 55–60 (2012).
82. Le Chatelier, E. et al. Richness of human gut microbiome correlates with metabolic markers. *Nature* **500**, 541–546 (2013).
83. Xie, H. et al. Shotgun metagenomics of 250 adult twins reveals genetic and environmental impacts on the gut microbiome. *Cell Syst.* **3**, 572–584 (2016).
84. Mitchell, A. L. et al. EBI Metagenomics in 2017: enriching the analysis of microbial communities, from sequence reads to assemblies. *Nucleic Acids Res.* **46**, D726–D735 (2018).
85. Yu, J. et al. Metagenomic analysis of faecal microbiome as a tool towards targeted non-invasive biomarkers for colorectal cancer. *Gut* **66**, 70–78 (2017).
86. Liu, W. et al. Unique features of ethnic mongolian gut microbiome revealed by metagenomic analysis. *Sci. Rep.* **6**, 34826 (2016).
87. He, Q. et al. Two distinct metacommunities characterize the gut microbiota in Crohn's disease patients. *Gigascience* **6**, 1–11 (2017).
88. Qin, N. et al. Alterations of the human gut microbiome in liver cirrhosis. *Nature* **513**, 59–64 (2014).
89. Human Microbiome Project Consortium. Structure, function and diversity of the healthy human microbiome. *Nature* **486**, 207–214 (2012).
90. Tang., G., Müller., M., Rios., A. & Sennrich., R. Why self-attention? A targeted evaluation of neural machine translation architectures. Preprint at https://arxiv.org/abs/1808.08946 (2018).
91. Vaswani., A. et al. Attention is all you need. Preprint at https://arxiv.org/abs/1706.03762 (2017).
92. Abadi, M. et al. TensorFlow: large-scale machine learning on heterogeneous distributed systems. Preprint at https://arxiv.org/abs/1603.04467 (2016).
93. Paszke., A. et al. PyTorch: an imperative style, high-performance deep learning library. *33rd Conference on Neural Information Processing Systems (NeurIPS 2019);* https://proceedings.neurips.cc/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf
94. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet.* **16**, 276–277 (2000).
95. Westbrook, A. et al. PALADIN: protein alignment for functional profiling whole metagenome shotgun data. *Bioinformatics* **33**, 1473–1478 (2017).
96. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
97. Segata, N. et al. Metagenomic microbial community profiling using unique clade-specific marker genes. *Nat. Methods* **9**, 811–814 (2012).
98. Zhang, B. & Horvath, S. A general framework for weighted gene co-expression network analysis. *Stat. Appl. Genet. Mol. Biol.* **4**, Article17 (2005).
99. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted correlation network analysis. *BMC Bioinformatics* **9**, 559 (2008).
100. Pollard, K. S., Gilbert, H. N., Ge, Y., Taylor, S. & Dudoit, S. multtest: resampling-based multiple hypothesis testing. *scienceopen.com* https://www.scienceopen.com/document?vid=43b5caa2-bac4-47c7-80d1-ee9c30ba9be7 (2011).
101. Shannon, P. et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
102. Cock, P. J. et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
103. Wayne, P. A. *Performance Standards for Antimicrobial Disk Susceptibility Tests* (Clinical and Laboratory Standards Institute, 1991).

Acknowledgements

This work was supported by the Strategic Priority Research Program of the Chinese Academy of Sciences (grant no. XDB29020000); the National Key Research and Development Program of China (grant nos. 2018YFC2000500 and 2018YFA0901900); the National Natural Science Foundation of China (grant nos. 32025002, 91857101 and 31771481); the Biological Resources Programme of the Chinese Academy of Sciences (grant no. KFJ-BRP-009); and the Beijing Nova Program (202077/202120).

Author contributions

J.W. and Y.C. conceptualized and managed this study. Y.M. developed the bioinformatics pipeline and screening. Y.M., X.L., B.X., Z.G., Y.Z., Y.Y., N.T., X.T. and M.W. carried out the experiments. Y.M., Z.G., B.X., X.L., X.Y., J.F., Y.C. and J.W. analyzed the data. Y.M., X.L., Y.C. and J.W. drafted the manuscript. J.F., Y.C. and J.W. edited the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

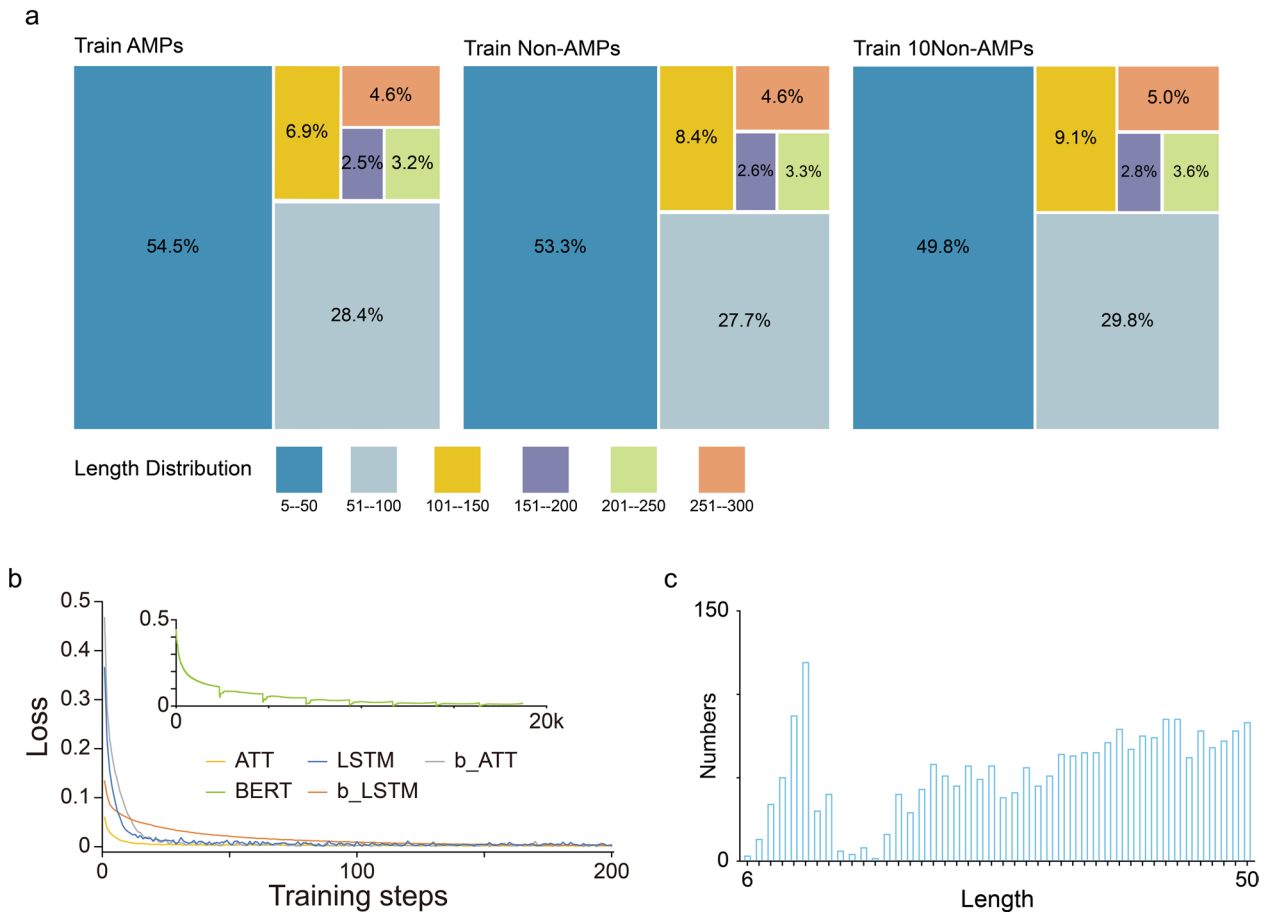
Extended data is available for this paper at https://doi.org/10.1038/s41587-022-01226-0.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41587-022-01226-0.

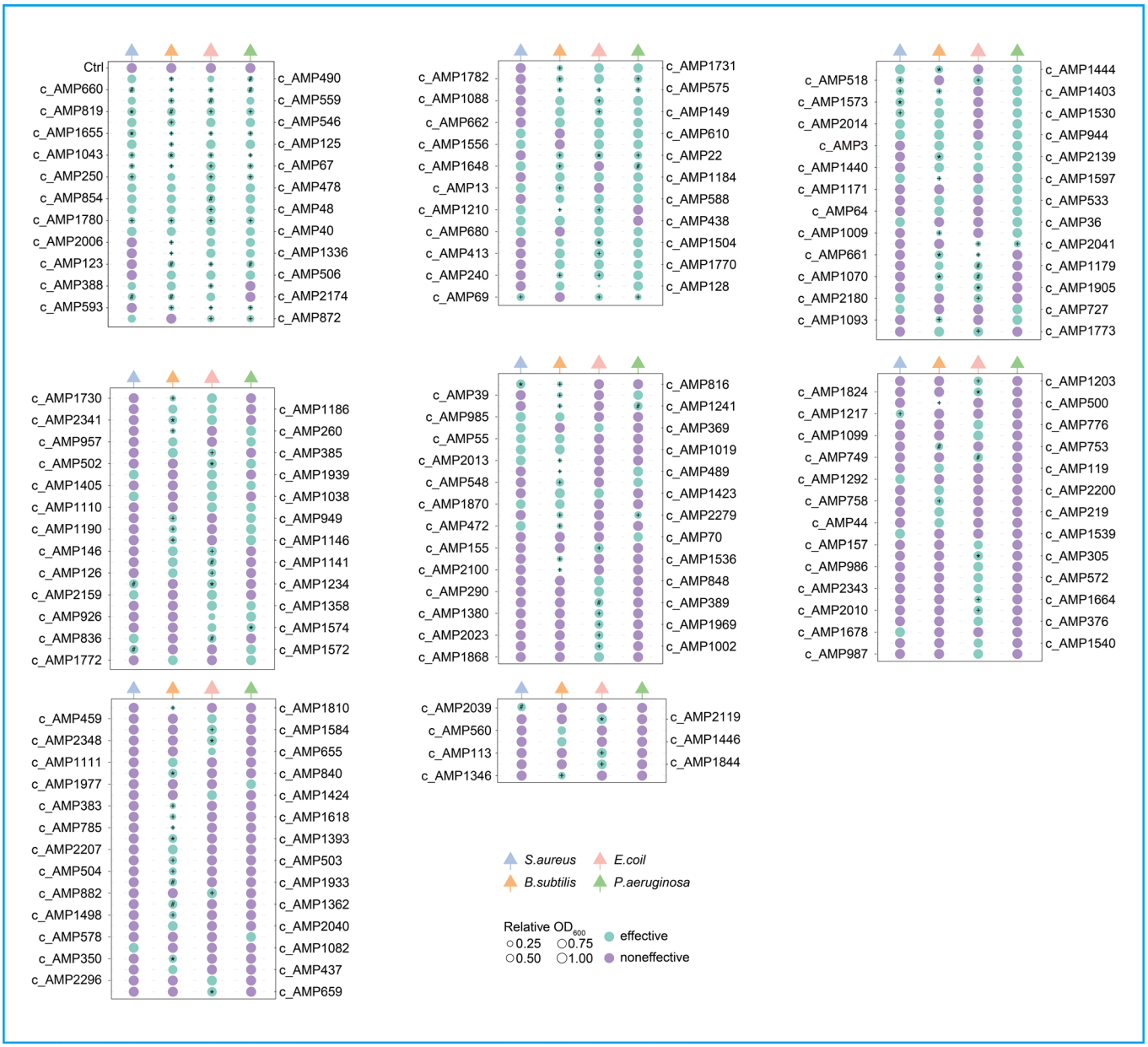
Correspondence and requests for materials should be addressed to Yihua Chen or Jun Wang.

Peer review information *Nature Biotechnology* thanks Luis Pedro Coelho and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

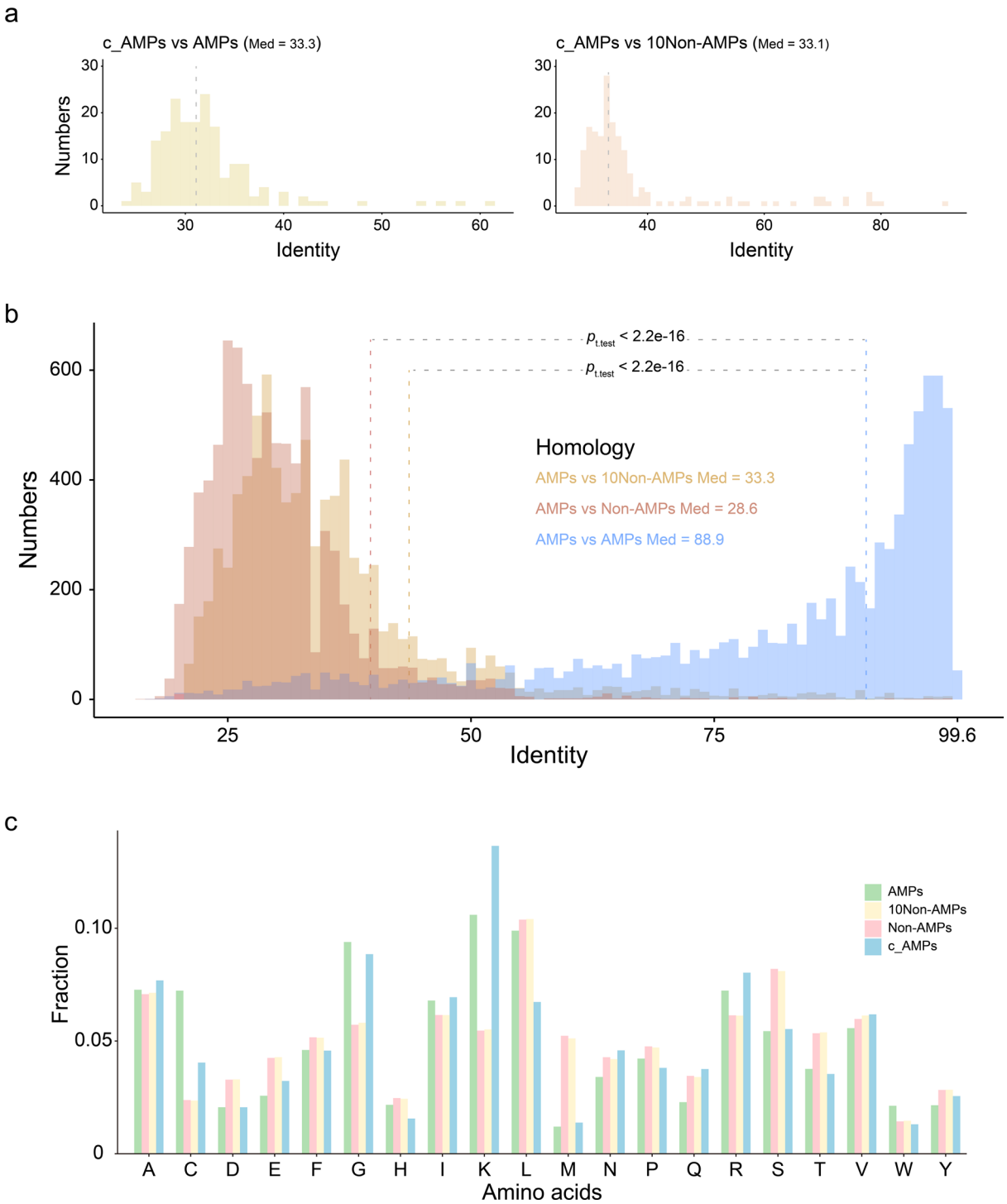
Reprints and permissions information is available at www.nature.com/reprints.



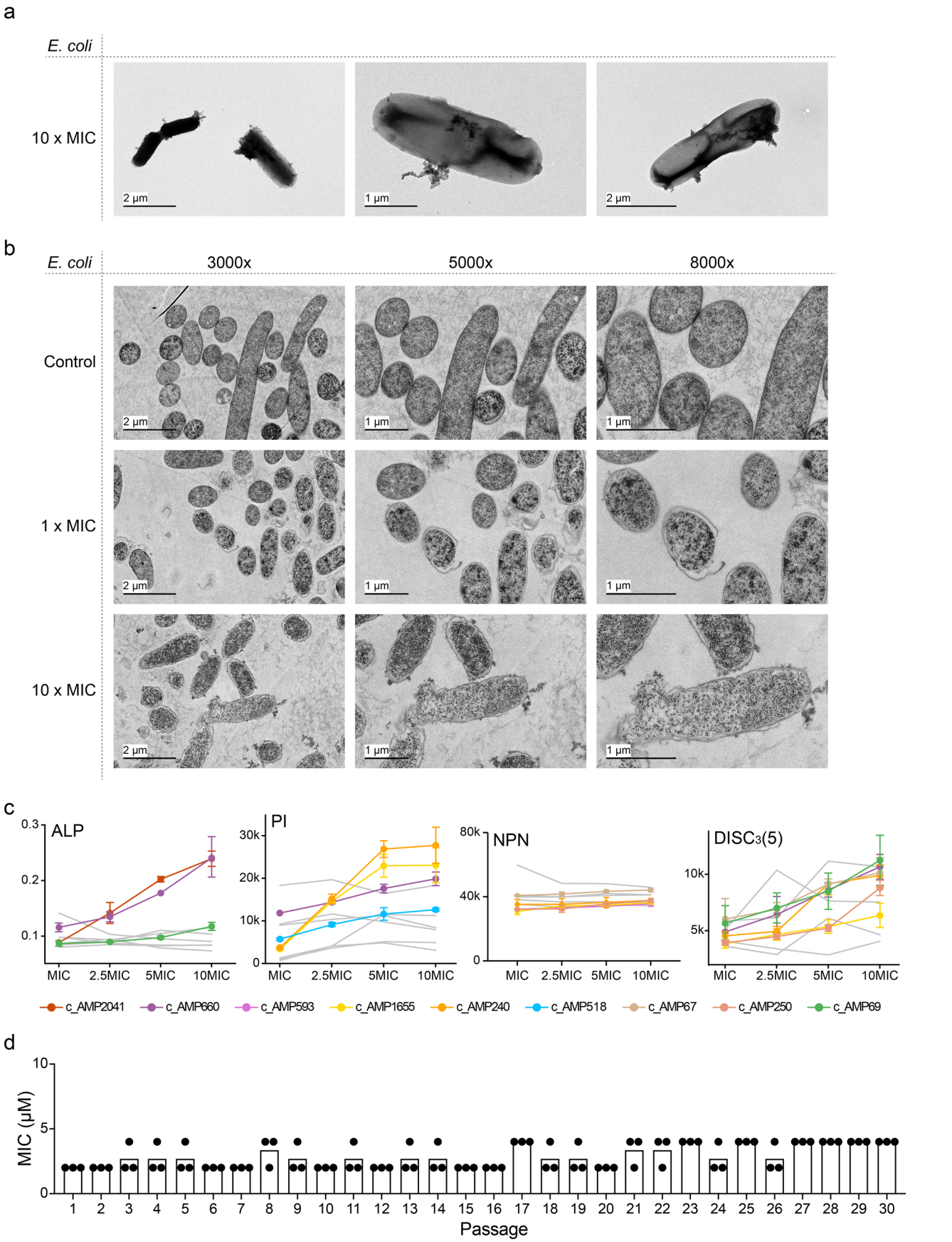
Extended Data Fig. 1 | Length distribution of datasets and model converge in the training stage. a, The length distribution of sequences in three training sets (Train AMPs: training set for AMP sequences, Train Non-AMPs: training set for non-AMP sequences with similar amount of sequences to that of Train AMPs, Train 10Non-AMPs: training set for non-AMP sequences with 10 times amount of sequences to that of Train AMPs) training data are matched. The colored squares indicate the different length distributions. This was plotted by http://www.bioinformatics.com.cn. b, The loss during training process of different models. Attention and LSTM models converged with 100-200 epochs of training steps, while Bert converged with higher number of epochs. c, Length distribution of 2,349 candidate AMPs from the metagenomic cohorts in our study.



Extended Data Fig. 2 | Spectra and level of bacterial inhibition of all c_AMPs against the four strains of bacteria used for the initial screening. Green color indicates that a c_AMPs significantly decreased the OD of at least one of the testing species. '*' denotes $0.01 < p \leq 0.05$, '#' denotes $0.001 < p \leq 0.01$ and '+' denotes $p \leq 0.001$, all in Dunnett's test (two-sided).

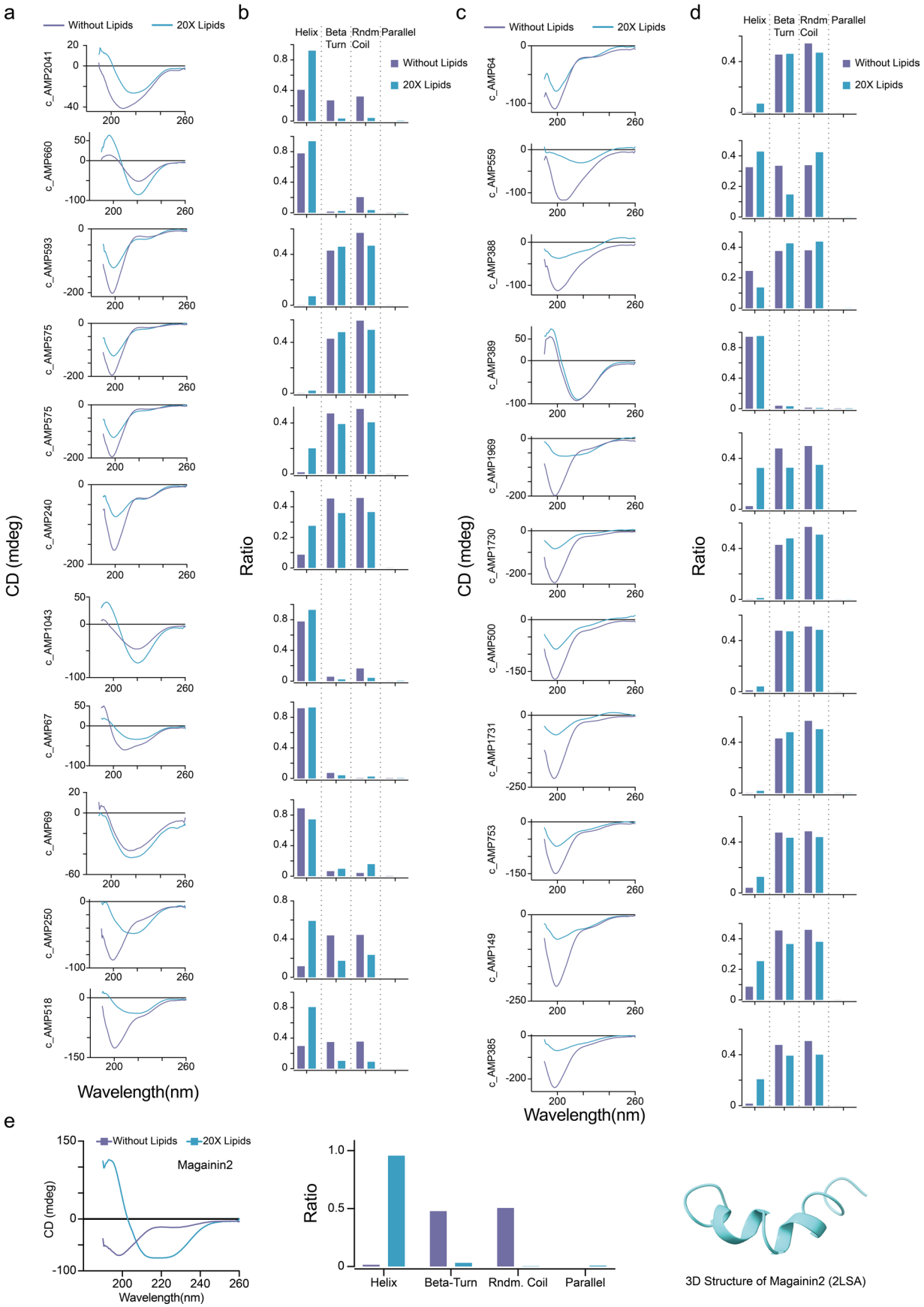


Extended Data Fig. 3 | Identities between AMPs and non-AMPs in our training set/discovered AMPs, and amino acid composition. a, Identity distributions based on multiple sequence alignment between c_AMPs and training set of AMPs/10Non-AMPs (see Methods), the grey line indicates the median of the identity values, Med stand for median identity. b, Sequence identity distributions in the training set, there was significantly higher identities among AMPs than between AMPs and non-AMPs (both balanced and unbalanced sets). One-sided Wilcoxon test was performed for each comparison. c, Amino acids composition of c_AMPs discovered in our study, and of known AMPs/non-AMPs in training sets (balanced dataset, pink; and unbalanced dataset, light yellow).



Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | c_AMPs mechanism of action and resistance development. a, Transmission electronic microscopy (TEM) examination of E. coli DH5α cells treated with c_AMP1043 at 10× MIC concentration, showing cell content leakage and cell wall/membrane disruption. Experiments were performed in triplicates with similar results and one representative figure is shown. b, Section photo of E. coli DH5α cells treated with c_AMP1043 and HEPES as control, with c_AMP1043 at MIC and 10× MIC concentration in the test, and three microscope images at difference magnifications were selected for each treatment. Experiments were performed in triplicates with similar results and one representative figure is shown. c, Mechanistic assays against E. coli DH5α for the ten other c_AMPs in the selected list. ALP, PI, NPN and DISC3(5) assays were used to examine the potential mechanism of function of c_AMPs, in particular the disruption of membrane of G- bacteria E. coli (see Methods and Results). Colored lines indicate dosage-dependent increase of signals. N = 3 independent experiments. Data are presented as mean values +/− SEM. d, Resistance development experiment of AMP1043 by serial passage against E. coli DH5α. The y-axis indicates the MIC measured directly from the tubes during the serial passages (μM) and the x-axis is the number of passages. In 30 passages, no observed resistance occurred as MIC remained <10 μM. N = 3 independent experiments.



Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Determination of peptide structures using circular dichroism (CD) spectra. a, CD results for the 11 most potent peptides and b, corresponding proportions of secondary structures calculated from CD data using CDNN. Purple: in water phase; dark blue: peptide mixed with 20 times of DMPE/DMPG lipid mixture (see Methods). c, and d, Further CD results and predicted structures of 11 randomly selected peptides with AMP activity. e, positive control Magainin 2, with CD results (left), predicted proportions of each secondary structure (middle) and known structure in PDB (right, accession no. 2LSA).

nature research

Corresponding author(s): Jun Wang, YiHua Chen

Last updated by author(s): Jan 13, 2022

Reporting Summary

Nature Research 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 Research 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	No software was used for data collection.
Data analysis	The c_AMPs prediction codes can be found at https://github.com/mayuefine/c_AMPs-prediction. In this study we used TensorFlow (version: 1.14.0) and PyTorch (version: 1.0.1) were used for NLP model training; R (version 3.4.1), Emboss software package (version 6.6.0.0), PALADIN (version 1.4.0),SAMtools (version 1.7) MetaPhlAn2 (MetaPhlAn 2.0), R package WGCNA (version 1.68), Cytoscape (version 3.6.1), Biopython module "Bio.SeqUtils.ProtParam" (version: 1.75), Prism (version 8.4.0)

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 Research 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 list of figures that have associated raw data
- A description of any restrictions on data availability

The analytical pipeline and original codes can be found at https://github.com/mayuefine/c_AMPs-prediction. Our study contains only publicly available AMP, non-AMP, metagenome and metaproteomics data, see Methods section for data collection. Our study contains only publicly available AMP, non-AMP, metagenome and metaproteomics data. AMP data were mainly collected from four public AMPs datasets, ADAM: http://bioinformatics.cs.ntou.edu.tw/adam/, APD: http://aps.unmc.edu/AP/main.php, CAMP: http://www.camp.bicnirrh.res.in/ and LAMP: http://biotechlab.fudan.edu.cn/database/lamp/, which cover most of AMP

sequences from different sources (downloaded as of 2018.10.02). The non-AMP dataset was downloaded from Uniprot (http://www.uniprot.org) by setting 'subcellular location' filter to cytoplasm and remove any entry that matches the following keywords: antimicrobial, antibiotic, antiviral, antifungal, effector or excreted (downloaded as of 2018.11.20). Validation datasets: Non-AMPs part ENA project ID is PRJEB19640, AMPs part were download from http://bagel4.molgenrug.nl/index.php. The representative genomes dataset was derived from Species-level genome bins: https://opendata.lifebit.ai/table/SGB. The metaproteome datasets were collected from https://www.ebi.ac.uk/pride, PRIDE project ID: PXD005780, PXD008870, PXD003907 and PXD000114. The 15 independent, large-scale metagenomic cohorts BioProject ID: PRJNA422434, PRJEB4336, PRJEB1220, PRJEB6337, PRJEB6456, PRJEB10878, PRJEB11532, PRJNA319574, PRJEB9584, PRJNA290380, PRJEB6337, PRJEB15371, PRJNA356102 and https://github.com/MetaSUB/MetaSUB-metadata.

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	Each group contained seven mice, the sample size was estimated based on a pilot study showing that seven mice are sufficient to observe significant differences between treatment and control groups.
Data exclusions	No data were excluded from the analysis
Replication	MIC and other determinations were replicated at least three times; all attempts at replications were successful.
Randomization	Mice were randomly allocated in each group receiving treatments.
Blinding	Blinding was not necessary for mice experiments.

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
☒	☐ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

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

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	HCT116 cells were purchased from ATCC
Authentication	Cells were authenticated using STR profiling by the vendor and no further authentication was performed in the laboratory.
Mycoplasma contamination	Tests for mycoplasma infection is routinely performed in the laboratory. All cells were negative for mycoplasma.
Commonly misidentified lines (See ICLAC register)	No such lines were used in this study.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	The experiment used 6 weeks old C57BL/6J mice, all female were used with 12/12 dark/light cycle, room temperature and humidity.

Wild animals	NA
Field-collected samples	NA
Ethics oversight	All animal experiments were approved by the Ethics Committee of the Institute of Microbiology, Chinese Academy of Sciences (SQIMCAS2021005).

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