ANTIMICROBIAL PEPTIDES

Antibiotic discovery with machine learning

Artificial intelligence finds candidate peptide antibiotics in the human gut microbiome.

Cesar de la Fuente-Nunez

Antimicrobial resistance has been highlighted as one of the top ten public health threats facing humanity by the World Health Organization. Antimicrobial peptides (AMPs) — small proteins typically 8–50 amino acids in length that confer protection against pathogens — are an established alternative to traditional antibiotics because they are less likely to elicit resistance[1]; however, only a limited number of these molecules have entered clinical practice, with dozens undergoing clinical and preclinical trials[1]. High-throughput approaches using microbiome data that widen the search for promising AMPs may provide a new source of candidates to combat antibiotic-resistant pathogens. Now Ma et al.[2] describe a clever artificial intelligence (AI) strategy to identify new antibiotics, employing natural language processing tools to effectively mine large gut microbiome datasets in search of peptides that possess antimicrobial properties (Fig. 1). The approach contributes to emerging research allowing the antibiotic discovery field to move past traditional methods that rely on arduous trial-and-error experimentation and into a new era where molecules can be rapidly discovered by computer.

Drug-resistant bacterial infections kill 1.27 million people worldwide each year[3,4], and without new classes of antimicrobial therapies, morbidity and mortality due to severe infections will increase: deaths caused by untreatable infections are projected to reach 10 million annually by 2050. The World Health Organization has highlighted five kinds of bacteria, called the ESKAPE pathogens, as priority pathogens that often display multi-drug resistance.

Several groups of researchers are using machine learning to discover new antibiotics[5–8], with approaches ranging from predictive to generative models. For example, generative models have been used to design novel AMPs with efficacy in animals[5] and displaying low toxicity[7]. Deep learning and other computational approaches have successfully repurposed molecules previously unrecognized to possess antibiotic activity[6] and discovered cryptic peptides with antimicrobial properties in the human body[8]. Finally, exciting comparative genomic pipelines have been developed to explore human microbiomes as a source of bioactive peptides and microproteins[9].

This work builds on these previous efforts by coupling computational tools with extensive experimental validation both in vitro and in relevant animal models. Ma et al. searched through large-scale metagenomics data, identifying sequences that resembled known AMPs. To devise a pipeline for AMP identification, the authors combined several deep-learning-based natural language processing models (for example, recurrent and attention neural networks), whose performance they optimized. As training data, the authors used extensive non-AMP datasets from the protein sequence database UniProt, increasing the dataset the model was trained on while minimizing the likelihood of false negatives. A total of 4,409 qualified representative genomes were searched, and 2,349 candidate AMPs were selected from expressed proteins with a length ranging from 6 to 50 amino acids. Next, Ma et al. investigated gene expression data, relative abundance and association with select bacterial taxa to remove unlikely AMPs from the list; this step led to the identification of 241 peptide sequences. Once these candidates were identified, the authors chemically synthesized the peptides and assessed their antimicrobial activity in vitro.

Of the 241 peptides, Ma et al. were able to synthesize 216, and 181 of these were found to have antimicrobial activity, resulting in a hit rate of 83.8%, superior to that in previous work, where 63.6% out of 55 synthesized peptides exhibited antimicrobial activity[8]. The authors then assessed the similarity of the 181 peptides with known AMP sequences present in the training set, revealing that the highest identity was

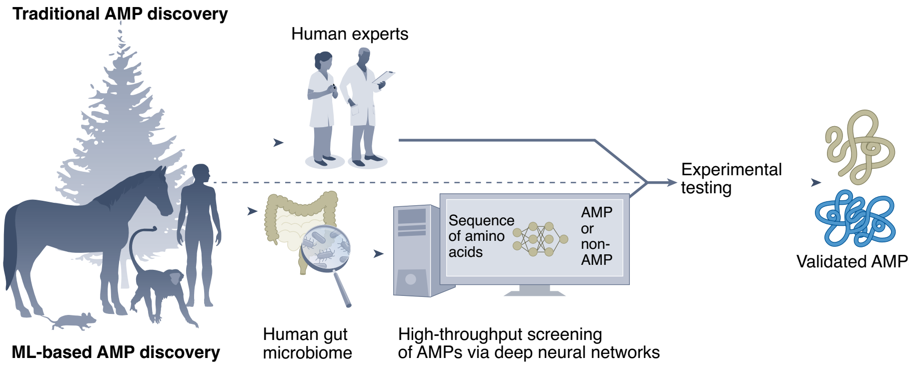

Fig. 1 | AI enables antibiotic discovery in the gut microbiome. A computational platform consisting of natural language processing and deep learning was used to explore human gut microbiome datasets searching for novel peptide antibiotics. ML, machine learning. (Image based on a design by Fangping Wan).

just 61.4%, with most sequences having identities of less than 40%. This analysis indicates that the peptides they discovered had sequences not previously associated with conventional AMPs.

One potential limitation of this and other computational approaches is their propensity toward biases. For example, a large number of peptide sequences present in UniProt start with a methionine amino acid, specified by the start codon AUG and likely irrelevant for bioactivity; this can bias the training since machine learning methods will build predictors with apparent (and overestimated) high accuracy taking into account methionine. Additional biases may further limit machine learning efforts aimed at drug discovery. However, Ma et al. synthesized and validated experimentally the antimicrobial activity of the peptides, demonstrating that, despite the potential for biases, their machine learning models were still able to discover AMPs effectively. Future work should focus on generating robust training sets and optimal models coupled with experimental validation for all or most sequences predicted and generated on the computer.

The 11 AMPs with the most potent antibacterial activity against drug-resistant bacteria, including ESKAPE pathogens, were selected for in-depth characterization. Among these 11 sequences, 7 were derived from *Bacteroides*, a dominant genus in the human gut microbiome, pointing to this genus as a potentially excellent source of AMPs. Peptide c_AMP1043 displayed the most potent antimicrobial activity, with a minimal inhibitory concentration of <10 μM against all clinical isolates tested, proving to be the most exciting lead candidate for follow-up studies. Mechanism-of-action studies with these 11 peptides indicated that the pipeline developed by Ma et al. may be able to capture AMPs with different mechanisms of action, even though this was not an input feature of the algorithms used. These results suggest that the computational methods used may reveal interesting hidden features within datasets.

A critical requirement when developing new drugs is that the targeted bacterial species do not develop resistance to the drug. To assess the potential development of bacterial resistance to c_AMP1043, the authors continuously exposed *Escherichia coli* strain DH5α to the peptide for 30 days, but detected no significant resistance. Critically, Ma et al. present in vivo data to establish the therapeutic potential of their strategy. Briefly, the authors show the low toxicity of the three lead peptides in cytotoxicity and hemolysis assays and validate the anti-infective efficacy of these agents against *Klebsiella pneumoniae* lung infections in a mouse model, showing that peptide treatment reduced bacterial load by >10-fold in vivo.

Collectively, Ma et al. present an AI approach based on natural language processing and deep learning to explore complex metagenomic information as a source of novel peptide antibiotics. Platforms such as the one described here are likely to transform antimicrobial research, making it possible to discover a greater variety of potential antibiotics in record time. We expect that the research by Ma et al. and other groups will effectively accelerate the discovery of new antibiotics that can be readily and easily adapted to treat severe bacterial infections. As this study demonstrates, AI approaches hold promise for the discovery of much-needed antimicrobial drugs, which can help to replenish our depleted arsenal.

Cesar de la Fuente-Nunez[1,2,3] ✉

[1]*Machine Biology Group, Departments of Psychiatry and Microbiology, Institute for Biomedical Informatics, Institute for Translational Medicine and Therapeutics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA.* [2]*Departments of Bioengineering and Chemical and Biomolecular Engineering, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA, USA.* [3]*Penn Institute for Computational Science, University of Pennsylvania, Philadelphia, PA, USA.*

✉*e-mail: cfuente@upenn.edu*

Published online: 6 May 2022
https://doi.org/10.1038/s41587-022-01327-w

References

1. Magana, M. et al. *Lancet Infect. Dis.* **20**, e216–e230 (2020).
2. Ma, Y. et al. *Nat. Biotechnol.* https://doi.org/10.1038/s41587-022-01226-0 (2022).
3. Der Torossian Torres, M. & de la Fuente-Nunez, C. *Chem. Commun. (Camb.)* **55**, 15020–15032 (2019).
4. Antimicrobial Resistance Collaborators. *Lancet* **399**, 629–655(2022).
5. Porto, W. F. et al. *Nat. Commun.* **9**, 1490 (2018).
6. Stokes, J. M. et al. *Cell* **181**, 475–483 (2020).
7. Das, P. et al. *Nat. Biomed. Eng.* **5**, 613–623 (2021).
8. Torres, M. D. T., Melo, M. C. R., Crescenzi, O., Notomista, E. & de la Fuente-Nunez, C. *Nat. Biomed. Eng.* **6**, 67–75 (2022).
9. Sberro, H. et al. *Cell* **178**, 1245–1259.e14 (2019).

Competing interests

The author declares no competing interests.

Check for updates

GENOMICS

Pathogenic or benign?

CRISPR base- and prime-editing pooled screens reveal the function of genetic variants at unprecedented resolution.

Peter P. Du, Katherine Liu, Michael C. Bassik and Gaelen T. Hess

High-throughput screens that rely on CRISPR tools to perturb gene expression are widely used to identify the biological functions of individual genes. Three new studies published in *Nature Biotechnology*[1–3] show how CRISPR base-editing and prime-editing technologies can be adapted to increase the resolution of these screens, enabling a more granular understanding of the effects of single point mutations. Systematic, genome-wide approaches such as these for functional testing of genetic variants are important not only in basic research but also in efforts to realize the promise of precision medicine.

Pooled screens that knock out genes with Cas9 nuclease or modulate gene expression with CRISPR inhibition or activation are widely used. They have identified many therapeutic targets and have defined the genes that regulate diverse cellular processes. However, these screens perturb expression of the entire gene product and thus do not capture the more nuanced effects of single point mutations or small in-frame deletions and insertions. To address this issue,