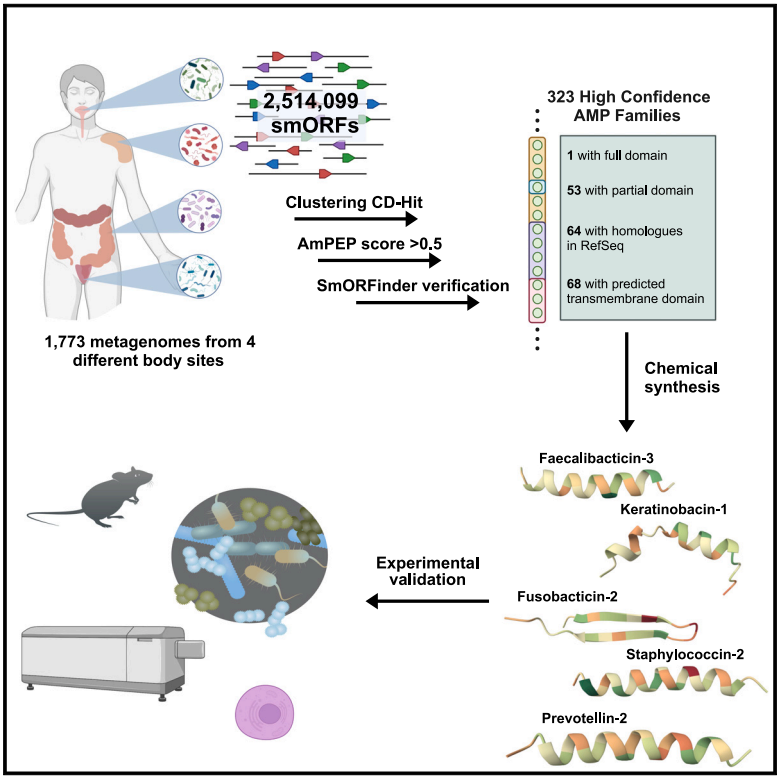
Cell

Mining human microbiomes reveals an untapped source of peptide antibiotics

Graphical abstract



Highlights

- Computational screen of human microbiome data identified candidate antimicrobials
- 323 smORF-encoded peptides with potential antimicrobial activity were identified
- 70.5% of synthesized smORF-encoded peptides showed antimicrobial activity *in vitro*
- Lead hit, prevotellin-2, has similar antibacterial efficacy to polymyxin B *in vivo*

Authors

Marcelo D.T. Torres, Erin F. Brooks, Angela Cesaro, ..., Cosmos Nicolaou, Ami S. Bhatt, Cesar de la Fuente-Nunez

Correspondence

asbhatt@stanford.edu (A.S.B.), cfuente@upenn.edu (C.d.l.F.-N.)

In brief

Computational analysis of 1,773 human gut metagenomes led to the discovery of 323 candidate antibiotics encoded in small open reading frames, known as smORF-encoded peptides (SEPs). These SEPs exhibit potent antibiotic properties against clinically relevant pathogens, both *in vitro* and *in vivo*.

Torres et al., 2024, Cell *187*, 5453–5467
September 19, 2024
https://doi.org/10.1016/j.cell.2024.07.027

CellPress

Cell

CellPress
OPEN ACCESS

Article

Mining human microbiomes reveals an untapped source of peptide antibiotics

Marcelo D.T. Torres,[1,2,3,4] Erin F. Brooks,[5] Angela Cesaro,[1,2,3,4] Hila Sberro,[5] Matthew O. Gill,[6] Cosmos Nicolaou,[5] Ami S. Bhatt,[5,6,*] and Cesar de la Fuente-Nunez[1,2,3,4,7,*]

[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 19104, USA
[2]Departments of Bioengineering and Chemical and Biomolecular Engineering, School of Engineering and Applied Science, University of Pennsylvania, Philadelphia, PA 19104, USA
[3]Penn Institute for Computational Science, University of Pennsylvania, Philadelphia, PA 19104, USA
[4]Department of Chemistry, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA 19104, USA
[5]Department of Medicine (Hematology; Blood and Marrow Transplantation), Stanford University, Stanford, CA 94305, USA
[6]Department of Genetics, Stanford University, Stanford, CA 94305, USA
[7]Lead contact
*Correspondence: asbhatt@stanford.edu (A.S.B.), cfuente@upenn.edu (C.d.l.F.-N.)
https://doi.org/10.1016/j.cell.2024.07.027

SUMMARY

Drug-resistant bacteria are outpacing traditional antibiotic discovery efforts. Here, we computationally screened 444,054 previously reported putative small protein families from 1,773 human metagenomes for antimicrobial properties, identifying 323 candidates encoded in small open reading frames (smORFs). To test our computational predictions, 78 peptides were synthesized and screened for antimicrobial activity *in vitro*, with 70.5% displaying antimicrobial activity. As these compounds were different compared with previously reported antimicrobial peptides, we termed them smORF-encoded peptides (SEPs). SEPs killed bacteria by targeting their membrane, synergizing with each other, and modulating gut commensals, indicating a potential role in reconfiguring microbiome communities in addition to counteracting pathogens. The lead candidates were anti-infective in both murine skin abscess and deep thigh infection models. Notably, prevotellin-2 from *Prevotella copri* presented activity comparable to the commonly used antibiotic polymyxin B. Our report supports the existence of hundreds of antimicrobials in the human microbiome amenable to clinical translation.

INTRODUCTION

The rapid emergence of antibiotic resistance in bacteria is a pressing threat to modern healthcare.[1–3] This has led to a dwindling supply of therapeutics to target multi-drug resistant pathogens, which are among the leading causes of nosocomial infections. Short peptides are a promising class of antimicrobial agents capable of addressing this threat.[3] Their short length of <50 amino acid residues, vast sequence space, and nonspecific mechanisms of action make them promising drug candidates that warrant exploration.[4]

Antimicrobial peptides (AMPs) can be encoded by unicellular and multicellular organisms. In metazoans, for example, AMPs are an ancient form of host defense.[4–6] In addition to directly targeting bacteria, certain "host-derived" peptides from metazoans have immunostimulatory properties that boost their potency and could be similarly leveraged by synthetic AMPs.[7] A growing number of AMPs derived from bacteria have also been described.[8,9] Interestingly, documented resistance to AMPs derived from prokaryotes or eukaryotes is exceedingly rare.[2,3] In those few cases where resistance has been documented, cross resistance to other AMPs[3] has not been demonstrated.

The physiochemical features of AMPs and the low rate of resistance to them have resulted in their use in clinical medicine. For example, widely used AMPs include bacitracin, colistin, and polymyxin B. Bacitracin, produced by *Bacillus licheniformis*, targets gram-positive bacteria by interfering with cell wall and peptidoglycan synthesis and is used to treat eye and skin infections. Polymyxin E (colistin), a last resort antibiotic produced by *Paenibacillus polymyxa* variant *collistinus*, has activity against gram-negative bacteria, working by displacing bivalent cations as counter ions of lipopolysaccharides that form the bacterial membrane. It is used mostly to treat pneumonia and biofilms in cystic fibrosis patients. Polymyxin B, produced by *Paenibacillus polymyxa*, an antibiotic used to treat topical infections and gut decontamination,[10] functions by disrupting the outer membrane of gram-negative bacteria. Thus, AMPs have potential for clinical utility; however, few other AMPs are commercially available. This is because identifying AMP candidates has been difficult. Current methods used to discover molecules rely primarily on prospection of naturally

occurring organisms, which is slow and unpredictable as it relies on trial-and-error experimentation. Additional engineering of AMPs has leveraged heuristics whereby amino acid residue substitutions are introduced,[11–15] guided by alterations in physicochemical feature determinants known to contribute to membrane targeting and, ultimately, antimicrobial activity.[4] Recently, advances in machine learning[16–18] and methods such as genetic algorithms[19,20] and pattern recognition algorithms[21,22] have yielded improved peptides. However, only a few efforts have focused on mining proteomes[23,24] and metagenomes.[23] The human microbiome, for example, offers significant promise in antimicrobial discovery but remains underexplored. For one, it has long been known that healthy members of the human microbiota can suppress the growth of pathogens. In addition, given the intense competition required to carve out a niche in this space, there is good reason to believe that microbial communities are significantly enriched for candidates with antimicrobial activity.

Although long overlooked due to the computational challenge of annotating genes encoding proteins $\leq$50 amino acid residues in length, the human microbiome was recently revealed to encode hundreds of thousands of small open reading frames (smORFs).[23] Of these, only a minuscule fraction has been functionally characterized. These represent a vast, untapped source of unexplored peptide sequences with potential antimicrobial activity. Indeed, bacteria have been shown to produce antimicrobial molecules to compete in complex ecosystems.[25] These molecules act through a variety of different mechanisms and can enable bacteria to kill related and unrelated strains, facilitating competition for limited niches.[26] For example, others have reported on an exciting class of anti-Bacteroidales peptide toxins from the gut microbiota.[25]

Recent research has leveraged natural language processing neural network models to explore the potential of discovering AMPs.[9] While this approach has successfully identified several potent peptides from stool metaproteomic datasets, it may have overlooked many peptides that are less abundantly expressed or originate from non-stool microbiome sources, such as the skin and oral cavity, which are known to contain a significant number of AMP genes.[8] In this work, we sought to take an unrestricted approach, considering small peptides derived from human metagenomes across a range of body sites. For this purpose, we started with a dataset of 444,054 predicted small peptides previously annotated from the Human Microbiome Project (HMP) metagenomes.[23,27] To narrow down this extensive list of candidate smORFs to one that is tractable for antimicrobial activity testing, we used existing computational algorithms to predict the likelihood that a given peptide sequence has antimicrobial activity, yielding a final list of 323 candidate smORF-encoded peptide (SEP) antibiotics. We then chemically synthesized 78 of these peptides and tested them against high-priority pathogens as well as against common gut commensals *in vitro*, finding 55 active peptides (70.5% hit rate). Five lead candidates from different sources, which had high activity against pathogens but limited or no activity against commensals, were tested *in vivo* and showed activity in preclinical infection animal models. Interestingly, analysis of data from public RNA sequencing (RNA-seq) and MetaRibo-seq studies[28] shows several candidates that are transcribed and translated, suggesting that at least some of these SEPs are produced in human microbiomes. Taken together, we leveraged our computational and experimental screening platform to mine a recently expanded microbiome microprotein database for candidate peptides with antimicrobial activity and identified highly potent and specific sequences (Figure 1).

RESULTS

In silico identification of candidate AMPs from the human microbiome

We first developed a discovery pipeline in which promising antimicrobial candidates were identified *in silico* from smORFs annotated in human metagenomes. We drew on a list of 444,054 families of putative small proteins (referred to as the 444,000 set) that were previously predicted using a comparative genomic workflow.[23] Briefly, MetaProdigal[29] was used to annotate all ORFs, as short as 15 base pairs, on 128,368,337 contigs spanning more than 180 billion base pairs of sequenced DNA from 1,773 metagenomes from 263 healthy individuals sampled from four distinct body sites. We eliminated ORFs that encoded proteins >50 amino acid residues in length and then clustered the remaining proteins based on sequence and length similarity using CD-Hit,[9] resulting in the 444,000 set. In this study, we mined the entire 444,000 set for candidate peptides with antimicrobial properties. We reasoned that those may be rapidly evolving and shared by fewer organisms,[30,31] in which case they might not be similar to other smORFs gene products in our dataset. Using AmPEP, a random forest classifier that predicts whether a given query sequence is likely to have antimicrobial activity or not through assigning a score between 0 and 1, we found 11,710 smORFs families that have an AmPEP score of $\geq$0.5, which indicates a given amino acid sequence is more likely than not to be an AMP.[11] To determine which of these smORFs are high-confidence protein-encoding genes, we identified those that were also identified by SmORFinder.[6] This tool combines profile hidden Markov models of each smORF family and deep learning models to predict smORF families that are likely to be valid. This reduced the list to 323 smORF-encoded candidate peptides with predicted antimicrobial activity (Data S1A), from which we selected 78 to synthesize and screen for biological activity. Selection of the final 78 peptides was based on four criteria: (1) high AmPEP score; (2) representation of the family of origin of the peptide (Data S1A), (e.g., clustered families with more members were preferentially selected over families with just one or a few candidate sequences); (3) amino acid composition that would enable an effective chemical synthesis, (e.g., sequences with motifs composed of amino acid residues with hindered side chains that would lead to low yield or many cysteine residues indicating constrained and complex secondary structures were filtered out); and (4) sequences without hydrophobic clusters were selected to avoid aggregation that would interfere with our subsequent screening effort.

Physicochemical features reveal a class of candidate peptide antibiotics

Physicochemical features provide valuable insights into the structural and molecular properties of AMPs, allowing for a

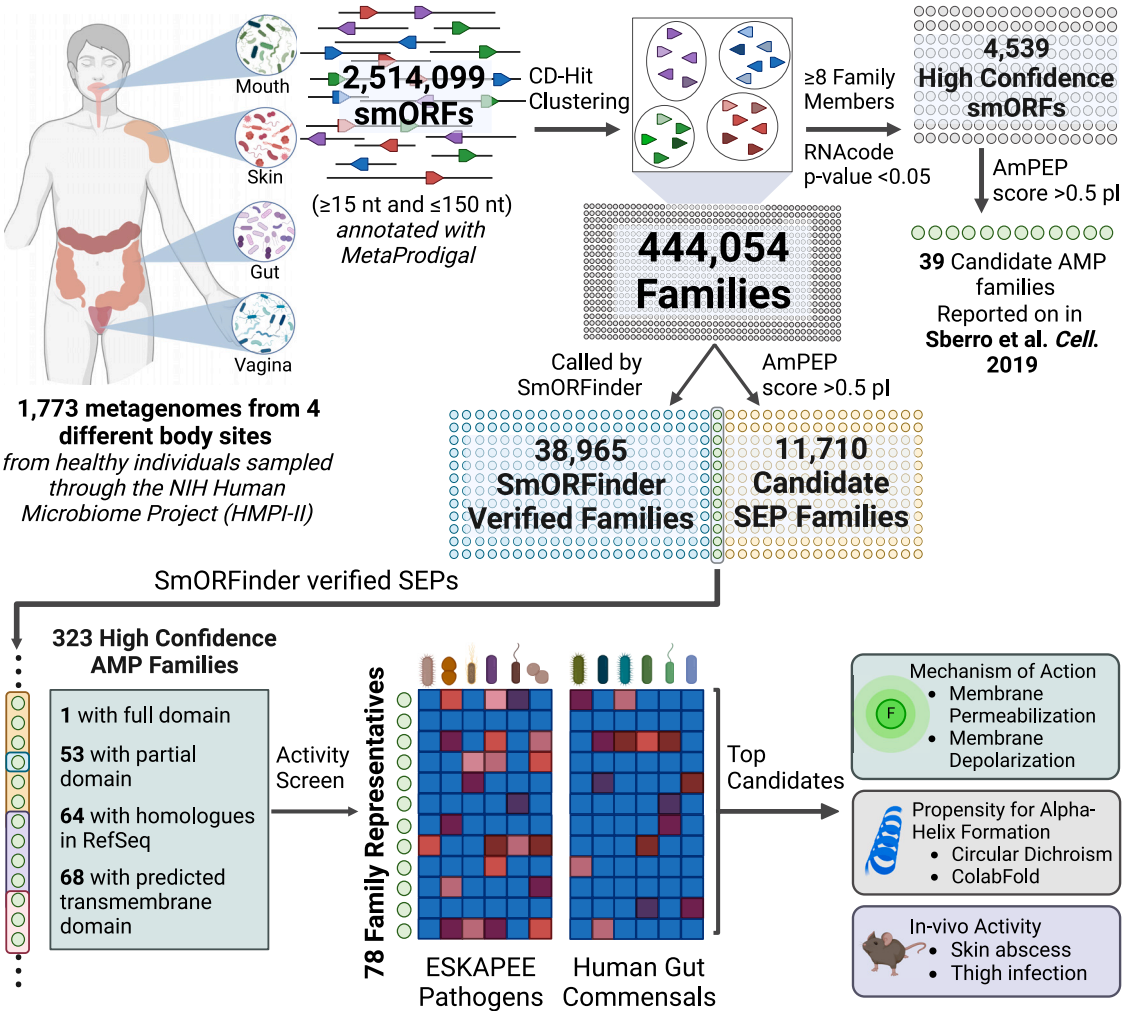


Figure 1. Schematic of the computational-experimental platform for the discovery of SEPs from smORFs

Metagenomes from four distinct body sites were analyzed to identify open reading frames (ORFs) containing more than 15 base pairs, using the MetaProdigal tool (see also Table S1). Subsequently, small ORFs (≤150 bp) were filtered out, and the encoded proteins were grouped into families, resulting in a total of 444,054 families. To further narrow down the selection, representatives of each family underwent analysis with SmORFinder, and the results were ranked using AmPEP to identify peptides with antimicrobial potential. The sequences that were identified by both SmORFinder and ranked as antimicrobials by AmPEP were considered as high-confidence families (see also Data S1A and S1B). These families were then subjected to further filtering based on specific criteria, as outlined in the inclusion and exclusion criteria for selecting peptides for activity testing section. The selected high-confidence families were subsequently tested against a range of pathogen and commensal bacterial strains. Promising candidates were further investigated through systematic characterization, including conformational studies, mechanism of action elucidation, assessment of synergistic interactions, and evaluation in preclinical mouse models. The figure was created with BioRender.com.

deeper understanding of their interactions with microbial membranes, their ability to penetrate cell membranes, and their stability in various biological environments. To assess the physicochemical features of all predicted (323 families) candidate smORF-encoded AMPs, we analyzed the amino acid composition (Figures 2A and S1) and known physicochemical determinants of antimicrobial activity (Figures 2B, 2C, and S2). We calculated their main physicochemical features using the Database of Antimicrobial Activity and Structure of Peptides (DBAASPs) server[32] (Figures 2B, 2C, and S2). For comparison purposes, we selected AMPs listed in DBAASP and encrypted peptides (EPs) found in the human proteome with predicted antimicrobial activity,[33] which are two classes of peptide antibiotics derived from different sources. Despite the fact that AmPEP predicts antimicrobial activity based on features found in known AMPs within its training set, the 323 SEP candidates displayed a higher content of negatively charged residues (aspartic and glutamic acids) than known AMPs and EPs (Figures 2A and S1A–S1D). This impacted their overall net charge, which is lower than most described AMPs and EPs (Figure 2B). The candidate SEPs also generally had a higher content of aliphatic and hydrophobic amino acids than hydrophilic ones (polar uncharged, acidic, and basic) compared with known AMPs and EPs (Figures 2A and S1B–S1D). Despite their high content of aliphatic

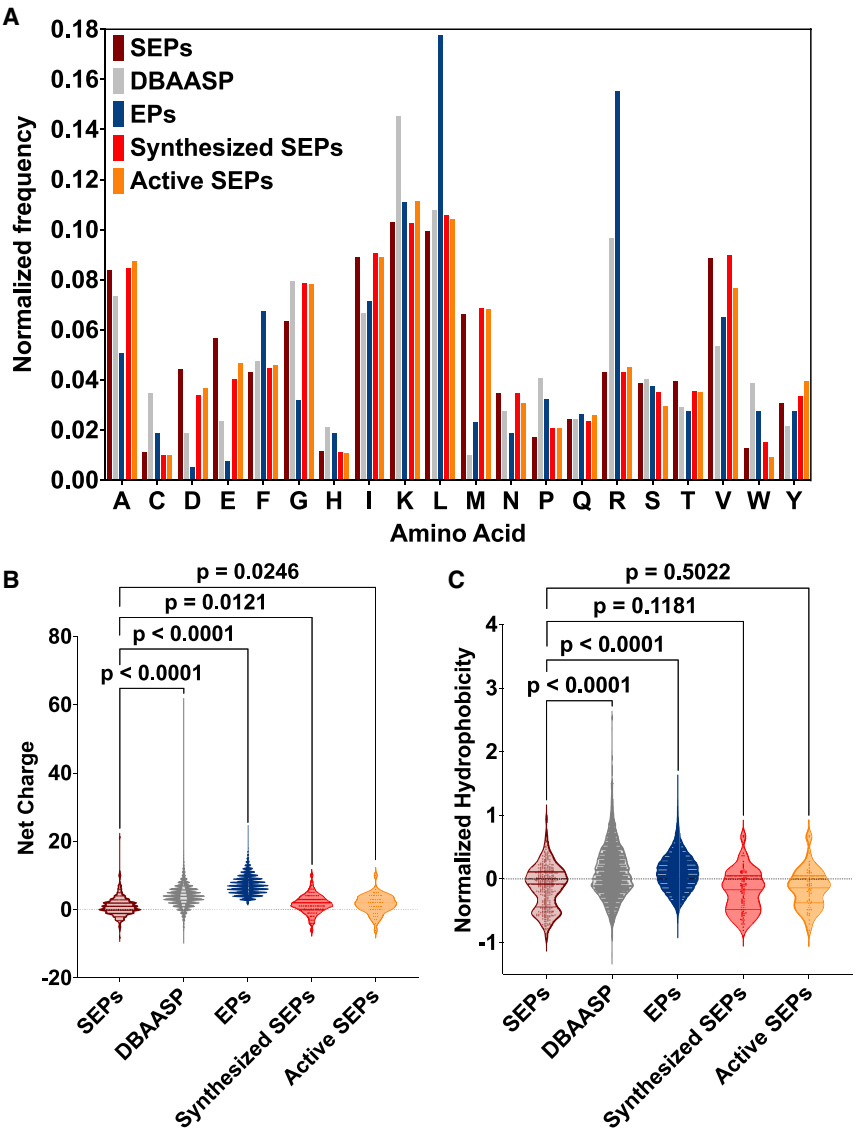


Figure 2. Sequence-related features of SEPs

(A) Amino acid frequency was calculated from candidate SEPs, from EPs predicted in the human proteome, and from AMPs in the DBAASP database. SEPs had overrepresentation of acidic residues (aspartic acid, D; and glutamic acid, E), polar uncharged residues (methionine, M, and asparagine, N), and lower content of leucine (L) residues. Synthesized and validated active SEPs showed similar amino acid content compared with all SEPs, and thus, they were considered as representatives of the total numbers of candidate SEPs. Among the most relevant physicochemical features that are known to influence biological activities of peptides (see also Figures S1 and S2), (B) SEPs have lower net positive charge and (C) normalized hydrophobicity than AMPs and EPs. Thus, SEPs are not amphipathic as other classes of AMPs, and instead, they are slightly less hydrophobic sequences with higher tendency to be disordered (see also Figure S2 and Table S2).

See also Figures S1 and S2.

and aromatic residues (Figures S1B–S1D), the low hydrophobic moment values (Figure S2A) summed to the high content of polar uncharged residues, such as methionine and asparagine, and to negatively charged residues make SEP candidates substantially less hydrophobic than previously described AMPs and EPs (Figures 2C, S1, and S2). The averaged lower hydrophobicity and net charge of the candidate SEPs directly influences their higher tendency to disordered conformation (Figure S2B), lower isoelectric point (Figure S2C), and lower amphiphilicity (Figure S2D) compared with already reported AMPs and EPs. The high content of methionine residues (Figure 2A) was expected since its associated codon, AUG, is the most common translation initiator.[34] Taken together, these physicochemical data reveal that these peptides do not rely on amphipathic structures with aliphatic and positively charged residues but instead consist of methionine-rich, slightly less hydrophobic sequences with an increased tendency to be disordered.

To verify that the SEPs identified and validated in this study represent a subset of the sequence space distinct from known AMPs, we compared their sequences against existing peptides in a public database, namely DBAASP. We initially conducted pairwise sequence similarity analyses between the 78 experimentally validated SEPs and peptides listed in this database. Subsequently, we applied the uniform manifold approximation and projection (UMAP) technique to transform these high-dimensional data into a two-dimensional (2D) space (Figure S2E). In this 2D UMAP visualization, while most known peptides formed clusters, many SEPs were positioned along the peripheries or in sparsely populated areas, indicating regions with few or no known peptides. Altogether, these data suggest that the peptides identified in this work constitute a class of sequences separate from known peptides such as AMPs and EPs.

Identification of SEP-encoding organisms in metagenomes and cultured isolates

For each SEP, we sought to learn more about its potential source organism. We evaluated contigs containing each of the original 323 predicted SEPs from our initial list by running a BLASTn search for each against NCBI's RefSeq nucleotide database with a significance E-value cutoff of <0. We chose the top hit organism based on the highest percent identity for every query as the putative source organism. In all cases, we found that the first and subsequently ranked hits all were to the same species (BLASTn results of the top hits are reported in Data S1B). Overall, of the 294 SEPs for which the genus of the source organism could be predicted, 69 different genera were represented. We

found that 29 SEPs were present on contigs without taxonomic hits or with hits to uncultured or unclassified organisms or plasmids. While many organisms on the list would be considered human commensals, including *Faecalibacterium*, *Prevotella*, *Bacteroides*, and *Lachnospiraceae* species, it was notable that several were known human pathogens or opportunistic pathogens, such as *Haemophilus parainfluenzae*, *Gemella haemolysans*, and *Escherichia coli* (*E. coli*). Interestingly, 78 (24%) of the 323 SEPs appeared to originate from viral or phage genomes, 13 of which belonged to the list of 78 SEPs chosen for testing. In addition, one SEP appeared to be from a human contig.

Given that we predicted our SEPs from metagenomic contigs, which precludes us from studying the organisms that make these SEPs experimentally, we were interested in identifying culturable bacterial strains that harbor these peptides in their genomes. We started by querying the genomes of each SEP's putative source organism but found no cases with culturable, readily available strain-level counterparts containing the exact respective antimicrobial smORF. We thus turned to *in silico* identification, this time using our existing set of 323 predicted SEPs to screen against proteins encoded in the recently published 119-species defined (gut) community, hCom2.[27] After predicting 531,822 ORFs as short as 15 base pairs in hCom2 genomes with MetaProdigal,[29] we screened the database for complete alignments (i.e., 100% amino acid identity across the entire length of the peptide) and found 17 SEPs aligning to 16 unique hCom2 organisms (see Table S1). Analysis of genomic neighborhoods revealed that 14/17 of these hCom2-positive SEPs overlap in-frame with existing gene annotations. While a few of these (3/14) share stop codons with the overlapping protein and could plausibly be translated separately from the encompassing coding sequence, the majority (11/14) share start codons with the overlapping protein with no subsequent logical stop codon. We reason that these homologous domains in hCom2 proteins, while not smORFs themselves, may be evolutionarily related to the SEPs that are detected in host-associated microbiota. Still, we found three SEPs without overlapping protein annotations that were identical to MetaProdigal-predicted smORFs in three hCom2 strains (*Bacteroides uniformis* ATCC 8492, *Bacteroides fragilis* 2-1-16, and *Bacteroides plebeius* DSM 17135) and were able to detect consistent expression of the *B. uniformis* ATCC 8492 SEP at the transcript-level in a public RNA-seq dataset of drug-treated isolates.[35] While these hits provide exciting impetus for future investigations of their SEP functions *in vitro*, the low detection rate (<1%) of our 323 predicted antimicrobial SEPs in hCom2, the most species-rich synthetic microbiota consortium to-date,[36] emphasizes a critical benefit of our approach: allowing insight into smORFs from un-isolated microbes.

Identification of transcription and translation of SEPs in published MetaRibo-seq data

To explore whether any of the SEPs that we identified might be expressed in microbial communities, we looked for evidence of their transcriptional and translational expression in paired, publicly available metatranscriptomic and MetaRibo-seq data from five human stool samples of varying levels of microbial diversity.[30] Five members of the list of 323 candidate SEPs were identified as transcribed and translated based on normalized read coverage of the SEP coding region in RNA-seq and MetaRibo-seq (Data S1C).[28] It is notable that two of these five (fusobacticin-1 and bacteroidin-1 putatively encoded by *Fusobacterium periodonticum* and *Bacteroides salanitronis*, respectively) were active with low MICs against several pathogens. Given that both SEPs are ostensibly encoded by gut commensals, evidence of their putative expression suggests they may play a key role in colonization resistance, especially against pathogenic strains of *E. coli* and *E. faecium*.

Antimicrobial activity of SEPs

To assess the potential antimicrobial activity of the candidate SEPs, we chemically synthesized 78 sequences (Table S2) and tested them against 11 clinically relevant pathogenic strains: *Acinetobacter baumannii* ATCC 19606, three *E. coli* strains (ATCC 11775, AIC221, AIC222) including a colistin-resistant strain, *Klebsiella pneumoniae* ATCC 13883, two *Pseudomonas aeruginosa* strains (PAO1 and PA14), two *Staphylococcus aureus* strains (ATCC 12600 and ATCC BAA-1556) including a methicillin-resistant strain, vancomycin-resistant *Enterococcus faecalis* (ATCC 700802), and vancomycin-resistant *Enterococcus faecium* (ATCC 700221), several of which are considered ESKAPEE pathogens (*E. faecium*, *S. aureus*, *K. pneumoniae*, *A. baumannii*, *P. aeruginosa*, *Enterobacter spp.*, and *E. coli*), and thus the most threatening bacterial pathogens in our society according to the World Health Organization (Figure 3A). Briefly, 10^6 cells mL^{-1} of each of the strains were exposed to a range of SEP concentrations (1–128 $\mu mol\ L^{-1}$). Growth (optical density at 600 nm, OD_{600}) was measured after incubating for 24 h at 37°C. Minimal inhibitory concentrations (MICs) values were determined as the concentrations where the peptide completely inhibits the growth of the bacteria (MIC_{100}). This initial screen yielded 33 SEPs (42.3% of the synthesized SEPs) that completely sterilized bacterial loads of at least one of the pathogens tested. The only strain not targeted by any of the SEPs was *S. aureus*.

Next, since SEPs were identified from human-associated metagenomes across several different body sites, including the gut, we screened them against 13 of the most abundant members of the human gut microbiota to assess whether they were able to target gut commensals.[37] The following bacteria were tested in anaerobic conditions: *Akkermansia muciniphila* ATCC BAA-635, *Bacteroides eggerthi* ATCC 27754, *Bacteroides fragilis* ATCC 25285, *Bacteroides ovatus* ATCC 8483, *Bacteroides thetaiotaomicron* ATCC 29148, *Bacteroides uniformis* ATCC 8492, *Bacteroides vulgatus* (*Phocaeicola vulgatus*) ATCC 8482, *Collinsella aerofaciens* ATCC 25986, *Clostridium scindens* ATCC 35704, *Clostridium spiroforme* ATCC 29900, *Eubacterium rectale* ATCC 33656, *Prevotella copri* DSM 18205, and *Parabacteroides distasonis* ATCC 8503 belonging to four different phyla (Verrucomicrobia, Bacteroidetes, Actinobacteria, and Firmicutes). Classical AMPs usually do not target microbiome strains[38]; however, we previously found that EPs were able to kill commensals.[33] Our screen yielded 45 SEPs with low micromolar antimicrobial activity against at least one of the gut commensals tested (57.7% hit rate; Figure 3A). We also found that all gut microbiome strains tested were susceptible to at least one SEP. In total, 55 of the 78 synthesized SEPs (70.5%)

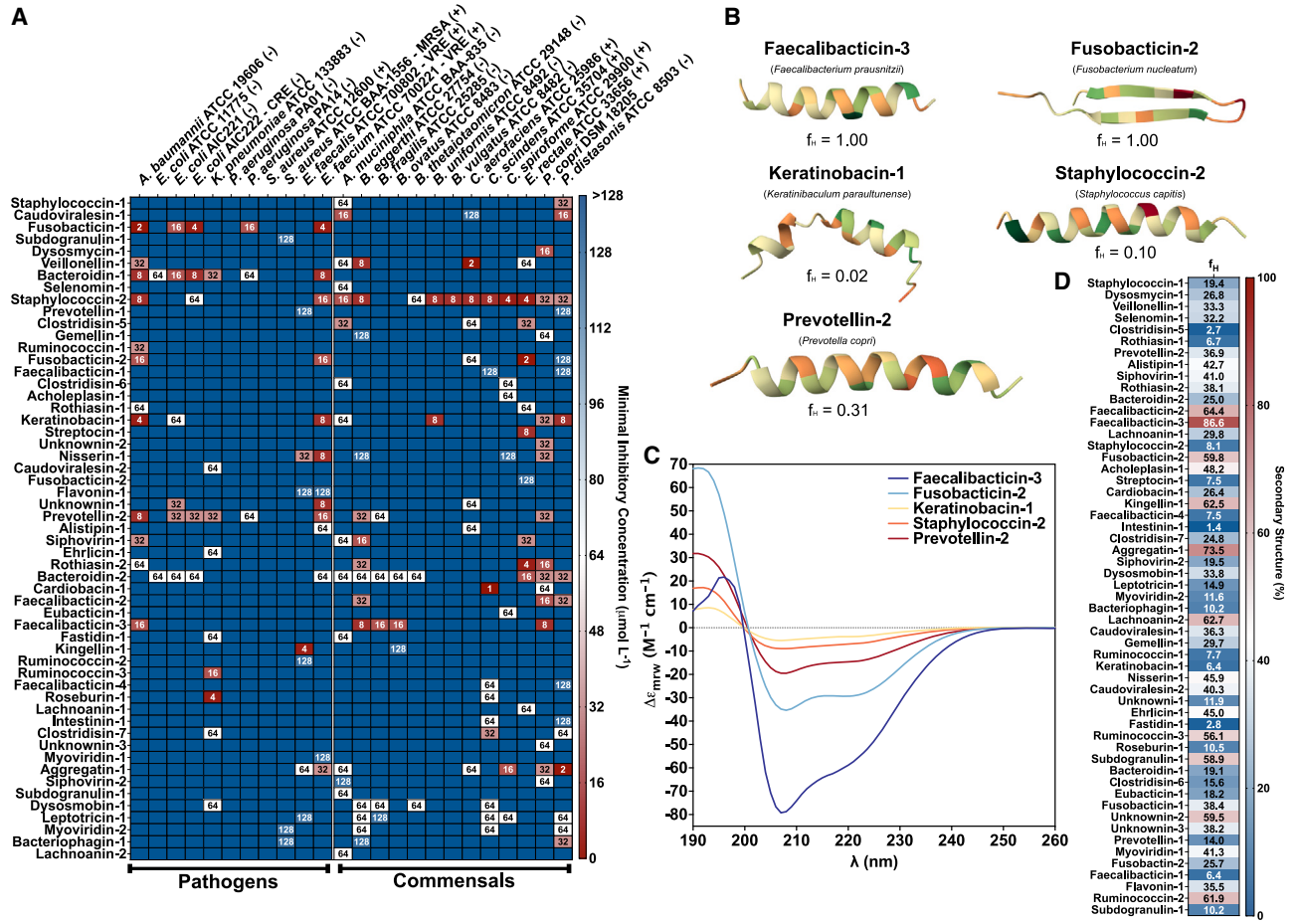


Figure 3. Antimicrobial activity and structure analysis of SEPs

(A) Antimicrobial activity of the tested SEPs. Briefly, a 10^6 bacterial cell load was exposed to serially diluted SEPs (1–128 μmol L^{-1}) in 96-well plates and incubated at 37°C. 20 h after the beginning of the experiment, each condition was analyzed in a microplate reader at 600 nm to check for inhibition of bacterial growth compared with the untreated controls. The results are presented as a heat map of antimicrobial activities (μmol L^{-1}) against pathogenic and gut commensal bacterial strains. Assays were performed in three independent replicates (see also Figure S2 and Data S1C).

(B) ColabFold was used to generate structural predictions using default parameters. Three-dimensional ribbon structures of the resulting PDB files were generated using Mol* 3D Viewer.

(C and D) (C) Circular dichroism spectra of the active SEPs in helical inducer medium to assess the tendency of SEPs to the most common structure adopted by antimicrobial peptides. Five of the SEPs present significant helical content (faecalibacticin-3, fusbacticin-2, keratinobacin-1, staphylococcin-2, and prevotellin-2) expressed in (D) helical fraction (f_H) values of active SEPs in a heat map, where the higher f_H values are presented in red and the lowest in blue. The activity did not correlate with the antimicrobial activity, once again reinforcing the independency of this class of peptides from amphipathic and balanced hydrophobic/cationic residues sequences (see also Figures S3A and S3B).

See also Figures S2 and S3.

had antimicrobial activity against at least one pathogen or commensal.

We found two cases where two SEPs were encoded by the same contig: ruminococcin-1 with ruminococcin-3 (*Ruminococcus bicirculans*) and caudoviralesin-1 with caudoviralesin-2 (*Caudovirales*). However, we found that no member of either of these pairs displayed notable activity against either pathogens or commensals, suggesting that they may fill another role, either individually or in tandem.

Staphylococcin-2 showed activity against the pathogens *A. baumannii* (8 μmol L^{-1}) and *E. faecium* (16 μmol L^{-1}) and displayed low MICs against nearly all commensals in our panel. Interestingly, it is encoded by a contig classified to *Staphylococcus capitis*, a member of the human skin microbiome. Closer examination of the contig (likely a plasmid) revealed the SEP was encoded 704 base pairs upstream from a type I toxin-antitoxin system Fst family toxin. A cursory search of other assemblies in NCBI containing this toxin revealed that it is encoded near staphylococcin-2 in several, but not all, assemblies examined.

Another interesting SEP was bacteroidin-2, which showed moderate activity (MIC values ranging from 16 to 64 μmol L^{-1}) against *E. coli*, vancomycin-resistant *E. faecium*, and many of the commensals in our panel. This SEP was encoded by a contig classified to *Bacteroides cellulosilyticus*, a common member of

the human gut microbiome known to ferment complex carbohydrates, including cellulose. It could be possible that *B. cellulosilyticus* uses bacteroidin-2 to maintain a niche within a healthy gut.

Investigation of prevotellin-2

Prevotellin-2, which is predicted to be encoded by a gut *P. copri* strain, was notable as one of the most potent antimicrobial SEPs tested in our study, showing an MIC of 8 μmol L^{-1} against *A. baumannii* and inactivity against most commensals except for *P. copri* DSM 18205, against which it displayed an MIC of 32 μmol L^{-1}. The activity that this SEP displays against another strain of the organism from which it is derived suggests that this SEP may mediate intraspecies antagonism.

Based on the observation that *P. copri* encodes a SEP that might enable intraspecies antagonism, we sought to determine whether this SEP is transcribed and translated. While the exact sequence of prevotellin-2 is not encoded in any *P. copri* reference strains, a homolog of prevotellin-2, differing by 3 amino acid residues, exists within *P. copri* DSM 18205. In investigating the genomic context of this prevotellin-2 homolog, we found that it is unique among our SEPs in that it is predicted to possess an alternate start codon (UUG, normally coding for leucine) and is encoded in-frame with the 3′ end of the 30S ribosomal protein S15 gene, *rpsO*. To determine if this prevotellin-2 homolog is transcribed, we performed RNA-seq on laboratory cultured *P. copri* DSM 18205. Briefly, we cultured *P. copri* DSM 18205 in rich media to exponential and early-stationary phases, isolated total RNA, and performed stranded RNA-seq. While we detected transcription of the prevotellin-2 homolog, in mapping read density across the *rpsO* gene, we did not detect differential read coverage in the 3′ end of the gene vs. the remainder of the transcript (see Figures S2F and S2G). This suggests that the prevotellin-2 homolog's transcript is not differentially expressed compared with the *rpsO* gene's transcript in these conditions. Taken together, we find that the prevotellin-2 homolog is transcribed, but does not appear to be differentially transcribed from the *rpsO* gene in which it is embedded. It is thus possible that in the type strain this SEP may or may not be transcribed and translated separately from the *rpsO* gene product.

Because prevotellin-2 and its homolog in *P. copri* DSM 18205 have 3 amino acids that differ, we sought to determine whether the DSM 18205 variant of prevotellin-2 possesses the same antimicrobial activity as the metagenome-derived SEP. This was especially important given that AMP activity can vary dramatically with minor changes in peptide sequence. Therefore, we synthesized two additional variants of prevotellin-2: prevotellin-2-1, which is identical to the 24 C-terminal amino acid residues of *rpsO* gene product S15 in *P. copri* DSM 18205, and prevotellin-2-2, which is identical to the 23 C-terminal amino acid residues of S15 but encodes a methionine in the first position of the sequence at the N-terminal extremity (based on the alternative start codon, UUG) rather than the leucine that is present in DSM 18205 (and prevotellin-2-1). Neither prevotellin-2-1 nor prevotellin-2-2 impacted the growth of ESKAPEE pathogens in our panel, despite the strong antimicrobial activity of the microbiome-encoded prevotellin-2 SEP (see Figure S2H). In summary, we find that prevotellin-2 is highly potent against pathogens, whereas the predicted homolog in the type strain of *P. copri* is much less potent and is not transcribed separately from the gene in which it is embedded in laboratory conditions. These findings highlight the subtle but important variations between related strains and the importance of mining metagenomes for antimicrobial smORFs, as their homologs in isolates, long-removed from host-associated communities, may be inactive.

Secondary structure of the active SEPs

The secondary structure of peptides dictates their antimicrobial and other biological activities. Since the synthesized SEPs were short (13–44 amino acid residues, with most containing ∼25 residues), they tended to be disordered in hydrophilic environments, such as water and buffers. To generate a first approximation of possible structure, we ran the 323 candidate SEP sequences through ColabFold using default parameters. ColabFold predicted all but two of the peptides to contain an α-helical domain. The two exceptions, SEPs fusobacticin-2 and roseburin-1, contained a β-like structure or were completely disordered, respectively. Although AMPs with high activity are often found to be α-helical, ColabFold predicted α-helical structures for both SEPs found to be active and those found to be inactive. Some of the predicted structures were interesting in that two distinct domains were predicted; an α-helical domain joined with a disordered region, which was the case for bacteroidin-2, clostridisin-5, and alistipin-1. One of the SEPs, staphylococcin-1, was notable for its kinked predicted α-helical structure. Five of the most active SEPs ColabFold structures are shown in Figure 3B. Given that AlphaFold is known to predict high-confidence α-helical structures for even spurious small proteins (<100 amino acid residues),[39] we probed the secondary structure tendencies of the peptides through circular dichroism in trifluoroethanol (TFE) and water mixtures (3:2, v:v). This solvent is a known α-helical inducer by dehydrating the amide groups that are part of the backbone of the peptide sequence and favoring intramolecular hydrogen bonds that lead the peptide to helical conformation.[40,41]

First, we took all SEPs at a fixed concentration (50 μmol L^{-1}) and obtained the circular dichroism spectra for wavelengths ranging from 260 to 190 nm (Figures 3C, S3A, and S3B). To obtain the secondary structure fractions, we used the Beta Structure Selection (BeStSel) server[42] (Figures 3D, S3C, and S3D). As expected, the ColabFold predicted secondary structures correlated poorly with the secondary structure fractions obtained experimentally (Figures S3C and S3D). Most of the active and inactive SEPs presented intermediary and low values of α-helical content in the conditions tested. Only ten of the active SEPs presented helical fraction (f_H) values higher than 50%. These results support the conclusion that SEPs tend toward a disordered conformation (Figures S3C and S3D) and that this does not adversely affect their antimicrobial activity.

Synergy between SEPs

Next, we wondered whether SEPs from the same biogeographical area of the body could synergize to target bacteria (Figure 4A). Checkerboard assays[33,43] were performed with pairs of SEPs derived from the same body sites, i.e., the tongue

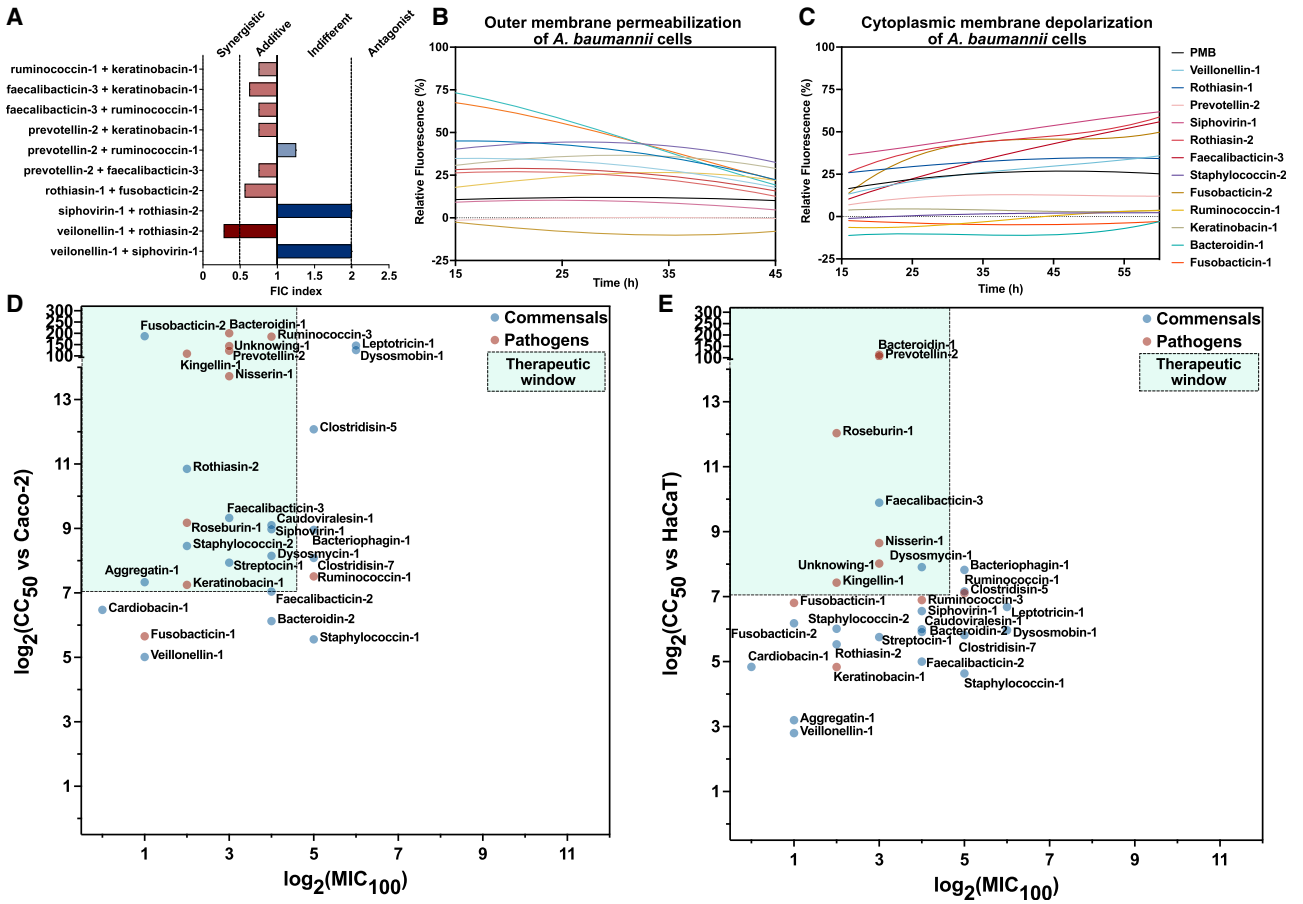


Figure 4. Synergy, mechanism of action, and cytotoxicity of SEPs

(A) The synergistic interaction between pairs of SEPs from the same biogeography (tongue dorsum, supragingival plaque, and stool) was assessed by checkerboard assays with 2-fold serial dilutions starting at 2 × MIC to MIC/32. The histogram shows the fractional inhibitory indexes (FICIs) values obtained for each pair of SEPs, where dark red represents synergistic interactions, light red indicates additive interactions, and blue shows indifferent interactions. Most of the pairs of SEPs presented synergistic or additive interactions. To assess whether SEPs act on the bacterial membrane, all active SEPs against each of the pathogenic strains were tested in outer membrane permeabilization and cytoplasmic membrane depolarization assays. In general, SEPs presented low permeabilization of the outer membrane effect, as shown in (B) the relative fluorescence measurements of SEPs on *A. baumannii* cell membranes (see also Figure S4). SEPs showed high depolarization properties as shown in (C) the relative fluorescence measurements of SEPs on vancomycin-resistant *E. faecium* cytoplasmic membranes (see also Figure S5). The relative fluorescence was calculated with a non-linear fitting using as baseline the untreated control (buffer + bacteria + fluorescent dye) as described in the STAR Methods section. The correlation between cytotoxicity on (D) human colorectal adenocarcinoma cells (Caco-2) or (E) immortalized human keratinocytes (HaCaT) and antimicrobial activity is shown in a scatterplot where the cytotoxicity is represented by the CC_{50} values (cytotoxic concentrations causing 50% cell death) and MIC (minimal inhibitory concentration for complete bacterial killing). CC_{50} values have been predicted by interpolating the dose-response with a non-linear regression curve. The green area represents the therapeutic window where those peptides could be safely used with no toxic effect to eukaryotic cells (see also Figures S5R and S5S).

See also Figures S4 and S5.

dorsum, supragingival plaque, and stool. To quantify the interactions between the SEP pairs, we extracted their fractional inhibitory concentration index (FICI).[44] Remarkably, almost all SEP pairs tested experimentally displayed synergistic or additive interactions against the gram-negative pathogen *A. baumannii*; the concentrations needed ranged from low micromolar to nanomolar concentrations *in vitro*, doses comparable to the most potent AMPs[12,14,45] and EPs.[33,46] One of the peptide pairs (veilonellin-1 and rothiasin-2) from the metagenome of bacteria from the tongue dorsum presented the most significant synergistic interaction with a FICI of 0.25.

Mechanism of action of SEPs

To investigate how SEPs exert their effects on bacterial cells, we conducted fluorescence assays to determine if they act by targeting the membrane. Firstly, we identified all antimicrobial hits among the SEPs (Figure 3A). Subsequently, we evaluated the capacity of these peptides (at their MIC value) to disrupt (Figures 4B and S4) and depolarize (Figures 4C and S5) the bacterial outer and cytoplasmic membrane, respectively. To assess whether SEPs can permeabilize the outer membrane of gram-negative bacteria, we performed 1-(N-phenylamino)naphthalene (NPN) assays. NPN is a lipophilic dye that emits fluorescence in

lipid-rich environments like bacterial outer membranes. If the bacterial outer membrane is damaged, NPN can permeate and increase its fluorescence (Figure 4B). The following gram-negative strains were exposed to SEPs in the presence of NPN: *A. baumannii* ATCC 19606 (Figure S4A), *E. coli* ATCC 11775 (Figure S4B), *E. coli* AIC221 (Figure S4C), *E. coli* AIC222 (Figure S4D), *K. pneumoniae* ATCC 13883 (Figure S4E), and *P. aeruginosa* PA14 (Figure S4F). All strains were permeabilized by the SEPs, except for *E. coli* ATCC 11775, which was permeabilized by all SEPs except for bacteroidin-2 (this peptide was only active in terms of antimicrobial activity against *E. coli* strains) (Figure 3A). Similarly, *P. aeruginosa* PA14 was not permeabilized by prevotellin-2, a broad-spectrum SEP that did permeabilize *A. baumannii* ATCC 19606, *E. coli* strains AIC221 and AIC222, and *K. pneumoniae* ATCC 13883. The peptide antibiotic polymyxin B was used as a positive control in our studies.[33] In summary, SEPs did not permeabilize the outer membrane of bacteria to the level previously reported for AMPs[11,45] or EPs,[33] indicating that the mechanism of action of SEPs might be independent of outer membrane permeabilization.

Next, we used 3,3′-dipropylthiadicarbocyanine iodide ($DiSC_3$-5), a fluorophore used to assess whether a compound depolarizes the bacterial cytoplasmic membrane. If there are imbalances in the transmembrane potential of the cytoplasmic membrane, the fluorophore migrates to the extracellular environment, producing fluorescence. Out of all 60 conditions tested, in 46 occasions, SEPs tested depolarized the cytoplasmic membrane of bacteria more substantially than groups treated with polymyxin B, a control that displays certain level of depolarization[33] (Figures S5A–S5I). Overall, these data suggest that SEPs operate preferentially via depolarizing the cytoplasmic membrane as opposed to permeabilizing the outer membrane, revealing a mechanism that is distinct from that of conventional AMPs[11,45] and EPs,[33] which tend to target the outer membrane.

Cytotoxicity assays

To assess the potential toxicity of SEPs toward mammalian cells, we tested the 29 SEPs with higher antimicrobial activity. The peptides were exposed to human colorectal adenocarcinoma cells (Caco-2) and immortalized human keratinocytes (HaCaT), which serve as models of intestine and skin epithelia, respectively.[47–49] In the case of Caco-2 cells, most peptides showed CC_{50} values (i.e., cytotoxic concentration that leads to 50% cell death) higher than 32 $\mu mol\ L^{-1}$ (Figures 4D and S5R). Nearly all sequences active against bacterial pathogens at low MIC values did not display toxic effects when tested on Caco-2 cells, except for peptides staphylococcin-1, veillonellin-1, and fusobacticin-1, which were toxic at 32, 32, and 64 $\mu mol\ L^{-1}$, respectively (Figure S5R). Interestingly, the predicted CC_{50} values of SEPs when exposed to human keratinocytes were lower than those obtained for colorectal cells (Figures 4E and S5S). Particularly, peptides staphylococcin-1, veillonellin-1, and aggregatin-1 displayed toxic effects when tested at 16, 8, and 4 $\mu mol\ L^{-1}$, respectively, toward HaCaT cells. The higher cytotoxic profile found for HaCaT rather than Caco-2 cells might be explained by the differences in lipid membrane composition. Nontumorigenic cells, such as HaCaT, present a lower content of negatively charged phospholipids exposed to the outer side of the membrane with respect to tumoral-originated cells such as Caco-2. The presence of negatively charged and polar uncharged residues (Figures 2A and 2B) in the SEPs' sequences hinders their electrostatic interactions with Caco-2 cell membranes.[50–52] Such interactions are known to play a crucial role as primary interactions between peptides and lipid bilayer and promote their approximation to the interface membrane-extracellular environment.[4] Keratinocytes were generally more susceptible than Caco-2 cells at relatively low concentrations of SEPs (4–8 $\mu mol\ L^{-1}$), although those values were higher than the values obtained in the antimicrobial assays (MICs). Therefore, to prioritize which SEPs to test in animal models, we delineated the therapeutic window of each sequence (Figures 4D and 4E) to ensure that the cytotoxic concentration of each peptide against both cell lines tested was at least 2-fold lower than their antimicrobial activity, i.e., $MIC \leq \frac{CC_{50}}{2}$.

Anti-infective activity of SEPs in two different preclinical animal infection models

To test if the lead SEPs retained their antimicrobial potency in complex living systems, we tested them in two mouse models, namely in mouse skin abscess[46,53,54] and deep thigh infection[24,33] models (Figure 5A). In both models, we used the pathogen *A. baumannii*, which causes infections in the blood, urinary tract, lungs, and topical wounds and is one the major causes of mortality in hospitalized patients due to high levels of antimicrobial resistance.[55] Five lead SEPs displayed potent activity against *A. baumannii*: prevotellin-2 (8 $\mu mol\ L^{-1}$), faecalibacticin-3 (16 $\mu mol\ L^{-1}$), staphylococcin-2 (8 $\mu mol\ L^{-1}$), fusobacticin-2 (16 $\mu mol\ L^{-1}$), and keratinobacin-1 (4 $\mu mol\ L^{-1}$), and thus were tested in both mouse models at their MIC value (Figure 3A).

The skin abscess infection was established with a bacterial load of 10^6 cells in 20 μL of *A. baumannii* onto a wounded area of the skin (Figure 5A). A single dose of each SEP at their respective MIC was delivered to the infected area. 2 days post-infection, prevotellin-2 markedly reduced the bacterial load by three orders of magnitude compared with the untreated control group. Its potency was comparable to the activity observed in the positive control group of mice treated with polymyxin B (Figure 5B). The other SEPs reduced the bacterial load by two orders of magnitude (Figure 5B). 4-days post-infection all SEPs and polymyxin B were still preventing bacterial growth, and prevotellin-2 and polymyxin B reduced the bacterial counts by three to four orders of magnitude compared with the untreated control. All the other SEPs reduced bacterial load by two to three orders of magnitude compared with the untreated control (Figure 5B). These results are promising since the SEPs were administered only once and after the establishment of the abscess, highlighting their anti-infective potential. Critically, no significant changes in weight, a proxy for toxicity, were observed in our experiments (Figure 5C).

Next, we assessed the efficacy of the same lead SEPs (prevotellin-2, faecalibacticin-3, staphylococcin-2, fusobacticin-2, and keratinobacin-1) in a murine deep thigh infection model (Figure 5A). This preclinical model is widely used to assess the antibiotic potential of compounds. Briefly, mice were administered two rounds of cyclophosphamide treatment for

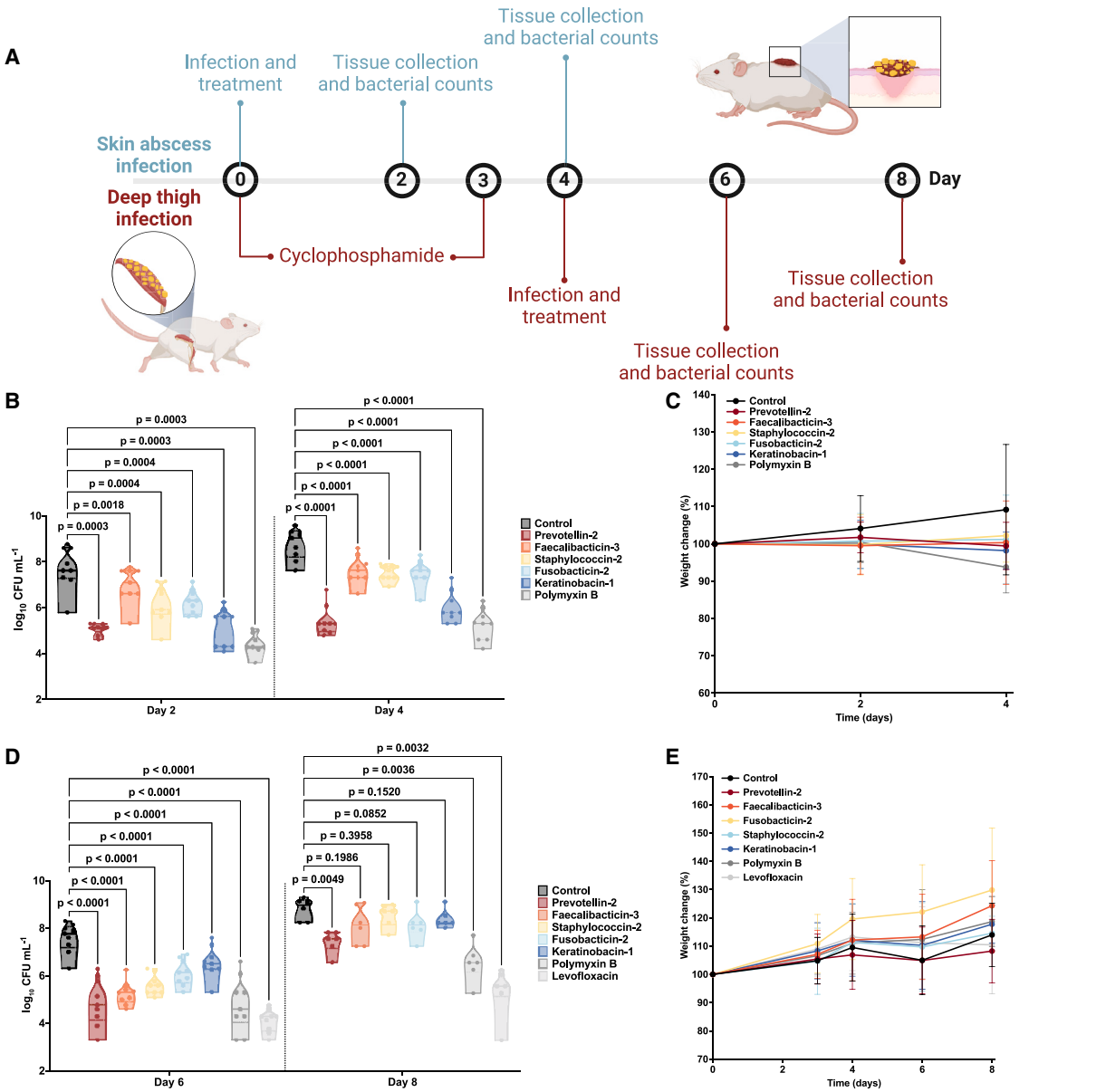


Figure 5. Anti-infective activity of SEPs in preclinical animal models

(A) Schematic of the skin abscess and deep thigh infection mouse models used to assess the anti-infective activity of the smORF-encoded peptides (SEPs) against *A. baumannii* cells.

(B) In the skin abscess infection model, mice were infected with a load of *A. baumannii* and treated 2 h after infection with one dose of the SEPs at their MIC. Mice were euthanized 2 and 4 days post-infection, and the tissue samples were homogenized and plated on agar plates in a 10-fold dilution gradient for colony-forming units (CFU) counts after an incubation of 24 h at 37°C. Each group (treated and untreated) consisted of three mice ($n = 3$), and the bacterial loads used for infection of each mouse came from a different inoculum. The experiment was done in three independent replicates. All peptides had similar bacteriostatic effect 2 days after infection, and after 4 days, all the SEPs tested were significantly different than the control, and the peptide prevotellin-2 presented activity comparable to the positive control (polymyxin B), reducing the infection by four orders of magnitude.

(C) To rule out toxic effects of the peptides, mouse weight was monitored throughout the whole extent of the experiment. We considered 20% weight change as acceptable considering the duration of the experiment.

(D) In the deep thigh infection mouse model, mice were first immunosuppressed by two rounds of treatment (24 and 72 h pre-infection) of immune system suppressor (cyclophosphamide). Subsequently, an intramuscular injection of *A. baumannii* was administered in the right thigh, followed by intraperitoneal administration of the peptides, to evaluate their systemic anti-infective activity 2 h after infection. 6 days after the start of the experiment, corresponding to 2 days post-infection, mice were euthanized. Each group, comprising treated and untreated mice, consisted of three individuals ($n = 3$), with distinct bacterial loads used for infecting each mouse, originating from different inocula. The experiment was done in three independent replicates. All peptides presented significant activity (one to two orders of magnitude reduction in bacterial counts), and the SEP prevotellin-2 had a similar effect than the antibiotics used as positive controls,

(legend continued on next page)

immunosuppression before the intramuscular infection with 10^6 cells in 100 μL of the bacterial pathogen *A. baumannii*. A single dose of each SEP (at their MIC) was delivered intraperitoneally (Figure 5A). 2 days post-treatment, prevotellin-2 and the antibiotics polymyxin B and levofloxacin (positive controls) reduced the bacterial load by three to four orders of magnitude (Figure 5D). All the other peptides led to a one to two orders of magnitude decrease in bacterial counts compared with the untreated control group of mice (Figure 5D). 4 days post-treatment, the bacterial counts increased for all peptide treatment conditions and the treatment with polymyxin B, while levofloxacin was still significantly active. In our experiments (Figure 5E), we did not observe any significant changes in weight, indicating that the SEPs are non-toxic. The *in vivo* results support the antibiotic properties of SEPs under physiological conditions and provide a strong basis for advancing their development as potential antimicrobial agents.

DISCUSSION

Here, we curated a list of 323 high-confidence SEP families predicted to be expressed in the human microbiome and holding great promise as antimicrobials. We synthesized and tested 78 of these and found that more than half displayed antimicrobial activity against at least one pathogen or commensal. The active SEPs were subjected to detailed characterization to determine their mechanism of action, secondary structure, and toxicity toward human cell lines. Interestingly, the five most promising SEPs were encoded by diverse phyla from oral, skin, and gut body sites: faecalibacticin-3 (*Faecalibacterium prausnitzii*), fusobacticin-2 (*Fusobacterium nucleatum*), keratinobacin-1 (*Keratinibaculum paraultunense*), staphylococcin-2 (*Staphylococcus capitis*), and prevotellin-2 (*Prevotella copri*). We tested these SEPs at their MIC to determine their *in vivo* anti-infective activity in skin abscess and deep thigh infection mouse models of *A. baumannii*. Our lead candidate, prevotellin-2, reduced bacterial loads at a comparable level to the current gold standard, polymyxin B, and without notable toxicity to the mammalian host in either infection model.

Upon analysis of relationships between producer and target organisms made apparent by our antimicrobial assays (Figure 3A), we observed three different general patterns of antagonism: (1) intraspecies antagonism, (2) interspecies within body-site antagonism, and (3) interspecies body-site exclusion. The first is the least surprising given the inherent threat posed by other strains occupying the same niche in the competition for resources. Narrow spectrum AMPs have been found to be produced by members of every major phylum of bacteria as well as in some Archaea and serve a crucial role in allowing producing strains to outcompete closely related bacteria with similar metabolic needs and lifestyle. These antimicrobials include the colicins of *E. coli*,[56] the pyocins of *Pseudomonas aeruginosa*,[57] the halocins of halobacteria, subtilin of *Bacillus subtilis*,[58] the lantibiotics of lactic acid bacteria (LAB), and other bacteriocins. In our dataset, the most notable example in this category is the antagonism observed between prevotellin-2 and *P. copri* DSM 18205 tested in our panel of commensals. It is interesting to note that the potential homolog of prevotellin-2 that is encoded in *P. copri* DSM 18205 has very low activity compared with prevotellin-2, suggesting that slight variations in SEP sequence can result in AMPs with highly varying activity. This supports the idea that SEPs are rapidly evolving and can vary in activity by strain. Our ability to observe further examples in this category is limited by the size of our panel.

Next, looking solely at our commensal activity panel, we observe two main categories of interspecies, broad-spectrum antagonism: within body-site antagonism and between body-site exclusion. The former, which was the most common in our dataset, included examples of varying ranges of target taxonomic specificity. Broader spectrum SEPs included several types of target class specificity. SEP faecalibacticin-3 from the gut microbiome (produced by *Faecalibacterium prausnitzii*, of the phylum Firmicutes) displayed cross-phylum antagonism targeting several specific organisms from the phylum Bacteroidetes. This pattern of phylum specific broad-spectrum antagonism is similar to the Bacteroidetes specific killing reported for bacteroidetocins encoded by certain Bacteroides species.[25,59] On the other hand, we also observe bacteroidin-2 (produced by *Bacteroides cellulosilyticus*), which targeted several different classes, including nearly all Bacteroidia in our panel (except for *B. uniformis* and *B. vulgatus*), as well as *Akkermansia muciniphila* and *Eubacterium rectale* of the Verrucomicrobiae and Clostridia classes, respectively.

Finally, we observe two examples of SEPs from other body sites targeting the gut commensals screened for sensitivity in our panel. The most striking example is staphylococcin-2 (produced by *Staphylococcus capitis*) from the skin microbiome, which displays broad activity across several phyla of gut commensals. In addition, we find it notable that fusobacticin-2 (produced by *Fusobacterium nucleatum*) from the oral microbiome has highly potent activity against the gut commensal *Eubacterium rectale*. We believe SEPs like these may play a role in shaping the niche of the producer organism by excluding non-native microbes. Future studies of the SEPs that are encoded in experimentally defined but complex communities, like hCom2, will help us shape our understanding of the native role of SEPs in modulating community composition.

Taken together, our results have important implications for how we think about determinants of microbiome community structure and the ability of invaders to establish a niche in a complex community. A growing body of evidence suggests that

polymyxin B and levofloxacin, reducing three to four orders of magnitude the bacterial counts 2 days post-infection (day 6 of the experiment). 4 days post-infection, levofloxacin was the only treatment that led to a more than 2 orders of magnitude decrease compared with the untreated control.
(E) During the entire 8-day period of the deep thigh infection model, mouse weight was closely monitored to eliminate the possibility of any toxic effects caused by cyclophosphamide injections, bacterial load, and the peptides. To determine statistical significance in (B) and (D), one-way ANOVA followed by Dunnett's test was employed, and the respective *p* values are presented for each group. All groups were compared with the untreated control, and the violin plots display the median and upper and lower quartiles. Data in (C) and (E) are the mean plus and minus the standard deviation. Figure created with BioRender.com.

microbes naturally produce antimicrobial molecules to compete in complex microbial communities.[25,60–63]

Our pipeline presents a platform for the discovery of peptide antibiotics, including those that might spare commensals. Here, we demonstrated the ability to find peptides that are as active as clinically relevant AMPs, such as polymyxin B. Future studies will help inform whether SEPs that are predicted to shape the microbiome are necessary and sufficient to do so in complex communities, and additional mining and drug development efforts are expected to identify SEPs that may have strong translational utility.

Limitations of the study

Although we find SEPs encoded by human-associated microbes with high levels of antimicrobial activity against both pathogens and commensals, our approach has limitations. First, we synthesized and tested the SEPs based on their translated coding sequences, which potentially overlooks those SEPs requiring post-translational modifications for full activity.[64] However, this may be of negligible concern given that all SEPs were computationally predicted to be antimicrobial as translated. This is supported by the high percentage of SEPs (70.5%) that showed activity in our screen. Second, given the limitations of Prodigal and SmORFinder, which we relied upon to identify and verify our list of SEPs, it is likely that we missed many SEPs with antimicrobial activity—both SEPs and those encoded within larger genes. Finally, while our results suggest varieties of intra- and interspecies warfare, our ability to draw conclusions about the role SEPs might play in bacterial competition and microbial community assembly is limited. While we have experimental evidence of transcription and translation for five members of the list, future metatranscriptomic and proteomic experiments will be informative in elucidating the *in vivo* expression for the majority of the SEPs. Moreover, the activity that we do observe could be incidental and not related to biological function. Looking ahead, as genomes and metagenomes become available, they promise to expand the pool of potential SEPs. All the SEPs predicted in this study were synthesized and experimentally validated in their linear form with no post-translational modifications.[64] Future work will aim to overcome these limitations by exploring SEPs requiring post-translational modifications and broadening our exploration into the SEP sequence space.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.cell.2024.07.027.

ACKNOWLEDGMENTS

C.d.l.F.-N. holds a Presidential Professorship at the University of Pennsylvania and acknowledges funding from the Procter & Gamble Company, United Therapeutics, a BBRF Young Investigator Grant, the Nemirovsky Prize, Penn Health-Tech Accelerator Award, Defense Threat Reduction Agency grants HDTRA11810041 and HDTRA1-23-1-0001, and the Dean's Innovation Fund from the Perelman School of Medicine at the University of Pennsylvania. Research reported in this publication was supported by the Langer Prize (AIChE Foundation), the NIH R35GM138201, and DTRA HDTRA1-21-1-0014. This work was also supported by a Paul Allen Distinguished Investigator Award and NIH R01AI148623 and R01AI143757 (to A.S.B.). We thank Dr. Mark Goulian for kindly donating the following strains: *Escherichia coli* AIC221 (*Escherichia coli* MG1655 phnE_2::FRT [control strain for AIC222]) and *Escherichia coli* AIC222 (*Escherichia coli* MG1655 pmrA53 phnE_2::FRT [polymyxin resistant]), Andrew Goodman for bacterial strains and assistance, and the H-MARC core in the Center for Molecular Studies in Digestive and Liver Diseases (P30 DK050306) for kindly donating strains *Acinetobacter baumannii* ATCC 19606 and *Klebsiella pneumoniae* ATCC 13883. We thank the de la Fuente lab and Bhatt lab members, especially Fangping Wan for assisting with the visualization of sequence space exploration and Soumaya Zlitni and Roby Bhattacharyya for insightful discussions. Figures created with BioRender.com are attributed as such.

AUTHOR CONTRIBUTIONS

Conceptualization, M.D.T.T., E.F.B., A.S.B., and C.d.l.F.-N.; methodology: M.D.T.T., E.F.B., A.C., H.S., A.S.B., C.N., and M.O.G.; investigation, M.D.T.T., E.F.B., A.C., and M.O.G.; visualization, M.D.T.T., E.F.B., A.C., A.S.B., and C.N.; funding acquisition, A.S.B. and C.d.l.F.-N.; supervision, A.S.B. and C.d.l.F.-N.; software, E.F.B., H.S., and C.N. formal analysis, M.D.T.T., E.F.B., and M.O.G.; writing – original draft, M.D.T.T., E.F.B., A.S.B., and C.d.l.F.-N.; writing – review & editing, M.D.T.T., E.F.B., A.C., C.N., H.S., M.O.G., A.S.B., and C.d.l.F.-N.

DECLARATION OF INTERESTS

C.d.l.F.-N. provides consulting services to Invaio Sciences and is a member of the Scientific Advisory Boards of Nowture S.L., Peptidus, and Phare Bio. De la

Fuente is also on the Advisory Board of the Peptide Drug Hunting Consortium (PDHC). The de la Fuente lab has received research funding or in-kind donations from United Therapeutics, Strata Manufacturing PJSC, and Procter & Gamble, none of which were used in support of this work. A.S.B. is on the scientific advisory board of Caribou Biosciences and Cantata Biosciences and is a scientific founder on the scientific advisory board and the Board of Directors for Stylus Medicine. An invention disclosure associated with the work has been submitted.

Received: August 8, 2023
Revised: May 9, 2024
Accepted: July 17, 2024
Published: August 19, 2024

REFERENCES

1. de la Fuente-Nunez, C., Torres, M.D., Mojica, F.J., and Lu, T.K. (2017). Next-generation precision antimicrobials: towards personalized treatment of infectious diseases. Curr. Opin. Microbiol. *37*, 95–102. https://doi.org/10.1016/j.mib.2017.05.014.
2. Mulani, M.S., Kamble, E.E., Kumkar, S.N., Tawre, M.S., and Pardesi, K.R. (2019). Emerging Strategies to Combat ESKAPE Pathogens in the Era of Antimicrobial Resistance: A Review. Front. Microbiol. *10*, 539. https://doi.org/10.3389/fmicb.2019.00539.
3. Magana, M., Pushpanathan, M., Santos, A.L., Leanse, L., Fernandez, M., Ioannidis, A., Giulianotti, M.A., Apidianakis, Y., Bradfute, S., Ferguson, A.L., et al. (2020). The value of antimicrobial peptides in the age of resistance. Lancet Infect. Dis. *20*, e216–e230. https://doi.org/10.1016/S1473-3099(20)30327-3.
4. Torres, M.D.T., Sothiselvam, S., Lu, T.K., and de la Fuente-Nunez, C. (2019). Peptide Design Principles for Antimicrobial Applications. J. Mol. Biol. *431*, 3547–3567. https://doi.org/10.1016/j.jmb.2018.12.015.
5. Wan, F., Wong, F., Collins, J.J., and de la Fuente-Nunez, C. (2024). Machine learning for antimicrobial peptide identification and design. Nat. Rev. Bioeng. *2*, 392–407. https://doi.org/10.1038/s44222-024-00152-x.
6. Wan, F., Torres, M.D.T., Peng, J., and de la Fuente-Nunez, C. (2024). Deep-learning-enabled antibiotic discovery through molecular de-extinction. Nat. Biomed. Eng. https://doi.org/10.1038/s41551-024-01201-x.
7. Hancock, R.E.W., and Sahl, H.-G. (2006). Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. Nat. Biotechnol. *24*, 1551–1557. https://doi.org/10.1038/nbt1267.
8. Santos-Júnior, C.D., Torres, M.D.T., Duan, Y., Rodríguez del Río, Á., Schmidt, T.S.B., Chong, H., Fullam, A., Kuhn, M., Zhu, C., Houseman, A., et al. (2024). Discovery of antimicrobial peptides in the global microbiome with machine learning. Cell *187*, 3761–3778.e16. https://doi.org/10.1016/j.cell.2024.05.013.
9. Ma, Y., Guo, Z., Xia, B., Zhang, Y., Liu, X., Yu, Y., Tang, N., Tong, X., Wang, M., Ye, X., et al. (2022). Identification of antimicrobial peptides from the human gut microbiome using deep learning. Nat. Biotechnol. *40*, 921–931. https://doi.org/10.1038/s41587-022-01226-0.
10. Severyn, C.J., Siranosian, B.A., Kong, S.T.-J., Moreno, A., Li, M.M., Chen, N., Duncan, C.N., Margossian, S.P., Lehmann, L.E., Sun, S., et al. (2022). Microbiota dynamics in a randomized trial of gut decontamination during allogeneic hematopoietic cell transplantation. JCI Insight *7*, e154344. https://doi.org/10.1172/jci.insight.154344.
11. Boaro, A., Ageitos, L., Torres, M.T., Blasco, E.B., Oztekin, S., and de la Fuente-Nunez, C. (2023). Structure-function-guided design of synthetic peptides with anti-infective activity derived from wasp venom. Cell Rep. Phys. Sci. *4*, 101459. https://doi.org/10.1016/j.xcrp.2023.101459.
12. Torres, M.D.T., Pedron, C.N., Higashikuni, Y., Kramer, R.M., Cardoso, M.H., Oshiro, K.G.N., Franco, O.L., Silva Junior, P.I., Silva, F.D., Oliveira Junior, V.X., et al. (2018). Structure-function-guided exploration of the antimicrobial peptide polybia-CP identifies activity determinants and generates synthetic therapeutic candidates. Commun. Biol. *1*, 221. https://doi.org/10.1038/s42003-018-0224-2.
13. Pedron, C.N., Torres, M.T., Lima, J.A.D.S., Silva, P.I., Silva, F.D., and Oliveira, V.X. (2017). Novel designed VmCT1 analogs with increased antimicrobial activity. Eur. J. Med. Chem. *126*, 456–463. https://doi.org/10.1016/j.ejmech.2016.11.040.
14. Torres, M.D.T., Pedron, C.N., Araújo, I., Silva, P.I., Silva, F.D., and Oliveira, V.X. (2017). Decoralin Analogs with Increased Resistance to Degradation and Lower Hemolytic Activity. ChemistrySelect *2*, 18–23. https://doi.org/10.1002/slct.201601590.
15. Torres, M.T., Pedron, C.N., da Silva Lima, J.A., da Silva, P.I., da Silva, F.D., and Oliveira, V.X. (2017). Antimicrobial activity of leucine-substituted decoralin analogs with lower hemolytic activity. J. Pept. Sci. *23*, 818–823. https://doi.org/10.1002/psc.3029.
16. Wan, F., Kontogiorgos-Heintz, D., and de la Fuente-Nunez, C. (2022). Deep generative models for peptide design. Digit. Discov. *1*, 195–208. https://doi.org/10.1039/D1DD00024A.
17. Wong, F., de la Fuente-Nunez, C., and Collins, J.J. (2023). Leveraging artificial intelligence in the fight against infectious diseases. Science *381*, 164–170. https://doi.org/10.1126/science.adh1114.
18. Cesaro, A., Bagheri, M., Torres, M., Wan, F., and de la Fuente-Nunez, C. (2023). Deep learning tools to accelerate antibiotic discovery. Expert Opin. Drug Discov. *18*, 1245–1257. https://doi.org/10.1080/17460441.2023.2250721.
19. Torres, M.T., and de la Fuente-Nunez, C. (2019). Toward computer-made artificial antibiotics. Curr. Opin. Microbiol. *51*, 30–38. https://doi.org/10.1016/j.mib.2019.03.004.
20. Porto, W.F., Irazazabal, L., Alves, E.S.F., Ribeiro, S.M., Matos, C.O., Pires, Á.S., Fensterseifer, I.C.M., Miranda, V.J., Haney, E.F., Humblot, V., et al. (2018). In silico optimization of a guava antimicrobial peptide enables combinatorial exploration for peptide design. Nat. Commun. *9*, 1490. https://doi.org/10.1038/s41467-018-03746-3.
21. Pizzo, E., Pane, K., Bosso, A., Landi, N., Ragucci, S., Russo, R., Gaglione, R., Torres, M.D.T., de la Fuente-Nunez, C., Arciello, A., et al. (2018). Novel bioactive peptides from PD-L1/2, a type 1 ribosome inactivating protein from Phytolacca dioica L. Evaluation of their antimicrobial properties and anti-biofilm activities. Biochim. Biophys. Acta Biomembr. *1860*, 1425–1435. https://doi.org/10.1016/j.bbamem.2018.04.010.
22. Pane, K., Cafaro, V., Avitabile, A., Torres, M.T., Vollaro, A., De Gregorio, E., Catania, M.R., Di Maro, A., Bosso, A., Gallo, G., et al. (2018). Identification of Novel Cryptic Multifunctional Antimicrobial Peptides from the Human Stomach Enabled by a Computational–Experimental Platform. ACS Synth. Biol. *7*, 2105–2115. https://doi.org/10.1021/acssynbio.8b00084.
23. Sberro, H., Fremin, B.J., Zlitni, S., Edfors, F., Greenfield, N., Snyder, M.P., Pavlopoulos, G.A., Kyrpides, N.C., and Bhatt, A.S. (2019). Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small, Novel. Genes. Cell *178*, 1245–1259.e14. https://doi.org/10.1016/j.cell.2019.07.016.
24. Maasch, J.R.M.A., Torres, M.D.T., Melo, M.C.R., and de la Fuente-Nunez, C. (2023). Molecular de-extinction of ancient antimicrobial peptides enabled by machine learning. Cell Host Microbe *31*, 1260–1274.e6. https://doi.org/10.1016/j.chom.2023.07.001.
25. Coyne, M.J., Béchon, N., Matano, L.M., McEneany, V.L., Chatzidaki-Livanis, M., and Comstock, L.E. (2019). A family of anti-Bacteroidales peptide toxins wide-spread in the human gut microbiota. Nat. Commun. *10*, 3460. https://doi.org/10.1038/s41467-019-11494-1.
26. Dobson, A., Cotter, P.D., Ross, R.P., and Hill, C. (2012). Bacteriocin Production: a Probiotic Trait? Appl. Environ. Microbiol. *78*, 1–6. https://doi.org/10.1128/AEM.05576-11.
27. Cheng, A.G., Ho, P.-Y., Aranda-Díaz, A., Jain, S., Yu, F.B., Meng, X., Wang, M., Iakiviak, M., Nagashima, K., Zhao, A., et al. (2022). Design, construction, and in vivo augmentation of a complex gut microbiome. Cell *185*, 3617–3636.e19. https://doi.org/10.1016/j.cell.2022.08.003.

28. Fremin, B.J., Sberro, H., and Bhatt, A.S. (2020). MetaRibo-Seq measures translation in microbiomes. Nat. Commun. *11*, 3268. https://doi.org/10.1038/s41467-020-17081-z.

29. Hyatt, D., Chen, G.-L., LoCascio, P.F., Land, M.L., Larimer, F.W., and Hauser, L.J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. BMC Bioinformatics *11*, 119. https://doi.org/10.1186/1471-2105-11-119.

30. Niehus, R., Oliveira, N.M., Li, A., Fletcher, A.G., and Foster, K.R. (2021). The evolution of strategy in bacterial warfare via the regulation of bacteriocins and antibiotics. eLife *10*, e69756. https://doi.org/10.7554/eLife.69756.

31. Smith, W.P.J., Wucher, B.R., Nadell, C.D., and Foster, K.R. (2023). Bacterial defences: mechanisms, evolution and antimicrobial resistance. Nat. Rev. Microbiol. *21*, 519–534. https://doi.org/10.1038/s41579-023-00877-3.

32. Pirtskhalava, M., Amstrong, A.A., Grigolava, M., Chubinidze, M., Alimbarashvili, E., Vishnepolsky, B., Gabrielian, A., Rosenthal, A., Hurt, D.E., and Tartakovsky, M. (2021). DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. Nucleic Acids Res. *49*, D288–D297. https://doi.org/10.1093/nar/gkaa991.

33. Torres, M.D.T., Melo, M.C.R., Flowers, L., Crescenzi, O., Notomista, E., and de la Fuente-Nunez, C. (2022). Mining for encrypted peptide antibiotics in the human proteome. Nat. Biomed. Eng. *6*, 67–75. https://doi.org/10.1038/s41551-021-00801-1.

34. Bhattacharyya, S., and Varshney, U. (2016). Evolution of initiator tRNAs and selection of methionine as the initiating amino acid. RNA Biol. *13*, 810–819. https://doi.org/10.1080/15476286.2016.1195943.

35. Ricaurte, D., Huang, Y., Sheth, R.U., Gelsinger, D.R., Kaufman, A., and Wang, H.H. (2024). High-throughput transcriptomics of 409 bacteria–drug pairs reveals drivers of gut microbiota perturbation. Nat. Microbiol. *9*, 561–575. https://doi.org/10.1038/s41564-023-01581-x.

36. van Leeuwen, P.T., Brul, S., Zhang, J., and Wortel, M.T. (2023). Synthetic microbial communities (SynComs) of the human gut: design, assembly, and applications. FEMS Microbiol. Rev. *47*, fuad012. https://doi.org/10.1093/femsre/fuad012.

37. Almeida, A., Mitchell, A.L., Boland, M., Forster, S.C., Gloor, G.B., Tarkowska, A., Lawley, T.D., and Finn, R.D. (2019). A new genomic blueprint of the human gut microbiota. Nature *568*, 499–504. https://doi.org/10.1038/s41586-019-0965-1.

38. Cullen, T.W., Schofield, W.B., Barry, N.A., Putnam, E.E., Rundell, E.A., Trent, M.S., Degnan, P.H., Booth, C.J., Yu, H., and Goodman, A.L. (2015). Gut microbiota. Antimicrobial peptide resistance mediates resilience of prominent gut commensals during inflammation. Science *347*, 170–175. https://doi.org/10.1126/science.1260580.

39. Monzon, V., Haft, D.H., and Bateman, A. (2022). Folding the unfoldable: using AlphaFold to explore spurious proteins. Bioinform. Adv. *2*, vbab043. https://doi.org/10.1093/bioadv/vbab043.

40. Luo, P., and Baldwin, R.L. (1997). Mechanism of helix induction by trifluoroethanol: A framework for extrapolating the helix-forming properties of peptides from trifluoroethanol/water mixtures back to water. Biochemistry *36*, 8413–8421. https://doi.org/10.1021/bi9707133.

41. Fioroni, M., Burger, K., Mark, A.E., and Roccatano, D. (2000). A new 2,2,2-trifluoroethanol model for molecular dynamics simulations. J. Phys. Chem. B *104*, 12347–12354. https://doi.org/10.1021/jp002115v.

42. Micsonai, A., Moussong, É., Wien, F., Boros, E., Vadászi, H., Murvai, N., Lee, Y.-H., Molnár, T., Réfrégiers, M., Goto, Y., et al. (2022). BeStSel: webserver for secondary structure and fold prediction for protein CD spectroscopy. Nucleic Acids Res. *50*, W90–W98. https://doi.org/10.1093/nar/gkac345.

43. Pletzer, D., Mansour, S.C., and Hancock, R.E.W. (2018). Synergy between conventional antibiotics and anti-biofilm peptides in a murine, sub-cutaneous abscess model caused by recalcitrant ESKAPE pathogens. PLoS Pathog. *14*, e1007084. https://doi.org/10.1371/journal.ppat.1007084.

44. Tyers, M., and Wright, G.D. (2019). Drug combinations: a strategy to extend the life of antibiotics in the 21st century. Nat. Rev. Microbiol. *17*, 141–155. https://doi.org/10.1038/s41579-018-0141-x.

45. Silva, O.N., Torres, M.D.T., Cao, J., Alves, E.S.F., Rodrigues, L.V., Resende, J.M., Lião, L.M., Porto, W.F., Fensterseifer, I.C.M., Lu, T.K., et al. (2020). Repurposing a peptide toxin from wasp venom into antiinfectives with dual antimicrobial and immunomodulatory properties. Proc. Natl. Acad. Sci. USA *117*, 26936–26945. https://doi.org/10.1073/pnas.2012379117.

46. Cesaro, A., Torres, M.D.T., Gaglione, R., Dell'Olmo, E., Di Girolamo, R., Bosso, A., Pizzo, E., Haagsman, H.P., Veldhuizen, E.J.A., de la Fuente-Nunez, C., and Arciello, A. (2022). Synthetic Antibiotic Derived from Sequences Encrypted in a Protein from Human Plasma. ACS Nano *16*, 1880–1895. https://doi.org/10.1021/acsnano.1c04496.

47. Colombo, I., Sangiovanni, E., Maggio, R., Mattozzi, C., Zava, S., Corbett, Y., Fumagalli, M., Carlino, C., Corsetto, P.A., Scaccabarozzi, D., et al. (2017). HaCaT Cells as a Reliable In Vitro Differentiation Model to Dissect the Inflammatory/Repair Response of Human Keratinocytes. Mediators Inflamm. *2017*, 7435621. https://doi.org/10.1155/2017/7435621.

48. Lea, T. (2015). Caco-2 Cell Line. In The Impact of Food Bioactives on Health: in vitro and ex vivo models, K. Verhoeckx, P. Cotter, I. López-Expósito, C. Kleiveland, T. Lea, A. Mackie, T. Requena, D. Swiatecka, and H. Wichers, eds. (Springer).

49. Boukamp, P., Petrussevska, R.T., Breitkreutz, D., Hornung, J., Markham, A., and Fusenig, N.E. (1988). Normal keratinization in a spontaneously immortalized aneuploid human keratinocyte cell line. J. Cell Biol. *106*, 761–771. https://doi.org/10.1083/jcb.106.3.761.

50. Arias, M., Haney, E.F., Hilchie, A.L., Corcoran, J.A., Hyndman, M.E., Hancock, R.E.W., and Vogel, H.J. (2020). Selective anticancer activity of synthetic peptides derived from the host defence peptide tritrpticin. Biochim. Biophys. Acta Biomembr. *1862*, 183228. https://doi.org/10.1016/j.bbamem.2020.183228.

51. Wang, C., Tian, L.-L., Li, S., Li, H.-B., Zhou, Y., Wang, H., Yang, Q.-Z., Ma, L.-J., and Shang, D.-J. (2013). Rapid cytotoxicity of antimicrobial peptide tempoprin-1CEa in breast cancer cells through membrane destruction and intracellular calcium mechanism. PLoS One *8*, e60462. https://doi.org/10.1371/journal.pone.0060462.

52. Szlasa, W., Zendran, I., Zalesińska, A., Tarek, M., and Kulbacka, J. (2020). Lipid composition of the cancer cell membrane. J. Bioenerg. Biomembr. *52*, 321–342. https://doi.org/10.1007/s10863-020-09846-4.

53. Silveira, G.G.O.S., Torres, M.D.T., Ribeiro, C.F.A., Meneguetti, B.T., Carvalho, C.M.E., de la Fuente-Nunez, C., Franco, O.L., and Cardoso, M.H. (2021). Antibiofilm Peptides: Relevant Preclinical Animal Infection Models and Translational Potential. ACS Pharmacol. Transl. Sci. *4*, 55–73. https://doi.org/10.1021/acsptsci.0c00191.

54. Arqué, X., Torres, M.D.T., Patiño, T., Boaro, A., Sánchez, S., and de la Fuente-Nunez, C. (2022). Autonomous Treatment of Bacterial Infections *in Vivo* Using Antimicrobial Micro- and Nanomotors. ACS Nano *16*, 7547–7558. https://doi.org/10.1021/acsnano.1c11013.

55. Karakonstantis, S., Gikas, A., Astrinaki, E., and Kritsotakis, E.I. (2020). Excess mortality due to pandrug-resistant Acinetobacter baumannii infections in hospitalized patients. J. Hosp. Infect. *106*, 447–453. https://doi.org/10.1016/j.jhin.2020.09.009.

56. Konisky, J. (1982). Colicins and other Bacteriocins with Established Modes of Action. Annu. Rev. Microbiol. *36*, 125–144. https://doi.org/10.1146/annurev.mi.36.100182.001013.

57. Michel-Briand, Y., and Baysse, C. (2002). The pyocins of Pseudomonas aeruginosa. Biochimie *84*, 499–510. https://doi.org/10.1016/S0300-9084(02)01422-0.

58. Schüller, F., Benz, R., and Sahl, H.G. (1989). The peptide antibiotic subtilin acts by formation of voltage-dependent multi-state pores in bacterial and

artificial membranes. Eur. J. Biochem. *182*, 181–186. https://doi.org/10.1111/j.1432-1033.1989.tb14815.x.

59. Matano, L.M., Coyne, M.J., García-Bayona, L., and Comstock, L.E. (2021). Bacteroidetocins Target the Essential Outer Membrane Protein BamA of *Bacteroidales* Symbionts and Pathogens. mBio *12*, e0228521. https://doi.org/10.1128/mBio.02285-21.
60. Roelofs, K.G., Coyne, M.J., Gentyala, R.R., Chatzidaki-Livanis, M., and Comstock, L.E. (2016). Bacteroidales Secreted Antimicrobial Proteins Target Surface Molecules Necessary for Gut Colonization and Mediate Competition In Vivo. mBio *7*, e01055-16. https://doi.org/10.1128/mBio.01055-16.
61. Evans, J.C., McEneany, V.L., Coyne, M.J., Caldwell, E.P., Sheahan, M.L., Von, S.S., Coyne, E.M., Tweten, R.K., and Comstock, L.E. (2022). A proteolytically activated antimicrobial toxin encoded on a mobile plasmid of Bacteroidales induces a protective response. Nat. Commun. *13*, 4258. https://doi.org/10.1038/s41467-022-31925-w.
62. Sugrue, I., Ross, R.P., and Hill, C. (2024). Bacteriocin diversity, function, discovery and application as antimicrobials. Nat. Rev. Microbiol. https://doi.org/10.1038/s41579-024-01045-x.
63. Chiumento, S., Roblin, C., Kieffer-Jaquinod, S., Tachon, S., Leprètre, C., Basset, C., Aditiyarini, D., Olleik, H., Nicoletti, C., Bornet, O., et al. (2019). Ruminococcin C, a promising antibiotic produced by a human gut symbiont. Sci. Adv. *5*, eaaw9969. https://doi.org/10.1126/sciadv.aaw9969.
64. Wang, G. (2012). Post-Translational Modifications of Natural Antimicrobial Peptides and Strategies for Peptide Engineering. Curr. Biotechnol. *1*, 72–79. https://doi.org/10.2174/2211550111201010072.
65. Fu, L., Niu, B., Zhu, Z., Wu, S., and Li, W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics *28*, 3150–3152. https://doi.org/10.1093/bioinformatics/bts565.
66. Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. *25*, 3389–3402. https://doi.org/10.1093/nar/25.17.3389.
67. Bhadra, P., Yan, J., Li, J., Fong, S., and Siu, S.W.I. (2018). AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. Sci. Rep. *8*, 1697. https://doi.org/10.1038/s41598-018-19752-w.
68. Durrant, M.G., and Bhatt, A.S. (2021). Automated Prediction and Annotation of Small Open Reading Frames in Microbial Genomes. Cell Host Microbe *29*, 121–131.e4. https://doi.org/10.1016/j.chom.2020.11.002.
69. Mirdita, M., Schütze, K., Moriwaki, Y., Heo, L., Ovchinnikov, S., and Steinegger, M. (2022). ColabFold: making protein folding accessible to all. Nat. Methods *19*, 679–682. https://doi.org/10.1038/s41592-022-01488-1.
70. Sehnal, D., Bittrich, S., Deshpande, M., Svobodová, R., Berka, K., Bazgier, V., Velankar, S., Burley, S.K., Koča, J., and Rose, A.S. (2021). Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. Nucleic Acids Res. *49*, W431–W437. https://doi.org/10.1093/nar/gkab314.
71. Li, H., and Durbin, R. (2009). Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics *25*, 1754–1760. https://doi.org/10.1093/bioinformatics/btp324.
72. Liao, Y., Smyth, G.K., and Shi, W. (2014). featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics *30*, 923–930. https://doi.org/10.1093/bioinformatics/btt656.
73. Lloyd-Price, J., Mahurkar, A., Rahnavard, G., Crabtree, J., Orvis, J., Hall, A.B., Brady, A., Creasy, H.H., McCracken, C., Giglio, M.G., et al. (2017). Strains, functions and dynamics in the expanded Human Microbiome Project. Nature *550*, 61–66. https://doi.org/10.1038/nature23889.
74. Zhao, M., Lee, W.-P., Garrison, E.P., and Marth, G.T. (2013). SSW Library: An SIMD Smith-Waterman C/C++ Library for Use in Genomic Applications. PLoS One *8*, e82138. https://doi.org/10.1371/journal.pone.0082138.
75. Prjibelski, A., Antipov, D., Meleshko, D., Lapidus, A., and Korobeynikov, A. (2020). Using SPAdes De Novo Assembler. Curr. Protoc. Bioinformatics *70*, e102. https://doi.org/10.1002/cpbi.102.
76. Schwengers, O., Jelonek, L., Dieckmann, M.A., Beyvers, S., Blom, J., and Goesmann, A. (2021). Bakta: rapid and standardized annotation of bacterial genomes via alignment-free sequence identification. Microb. Genom. *7*, 000685. https://doi.org/10.1099/mgen.0.000685.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Bacterial strains		
Acinetobacter baumannii	American Type Culture Collection	ATCC 19606
Escherichia coli	American Type Culture Collection	ATCC 11775
Escherichia coli	*Escherichia coli* MG1655 phnE_2::FRT	AIC221
Escherichia coli	*Escherichia coli* MG1655 pmrA53 phnE_2::FRT (polymyxin-resistant; colistin-resistant strain)	AIC222
Klebsiella pneumoniae	American Type Culture Collection	ATCC 13883
Pseudomonas aeruginosa	N/A	PAO1
Pseudomonas aeruginosa	N/A	PA14
Staphylococcus aureus	American Type Culture Collection	ATCC 12600
Staphylococcus aureus	American Type Culture Collection	ATCC BAA-1556 (methicillin-resistant strain)
Enterococcus faecalis	American Type Culture Collection	ATCC 700802 (vancomycin-resistant strain)
Enterococcus faecium	American Type Culture Collection	ATCC 700221 (vancomycin-resistant strain)
Akkermansia muciniphila	American Type Culture Collection	ATCC BAA-635
Bacteroides eggerthi	American Type Culture Collection	ATCC 27754
Bacteroides fragilis	American Type Culture Collection	ATCC 25285
Bacteroides ovatus	American Type Culture Collection	ATCC 8483
Bacteroides thetaiotaomicron	American Type Culture Collection	ATCC 29148
Bacteroides uniformis	American Type Culture Collection	ATCC 8492
Bacteroides vulgatus (*Phocaeicola vulgatus*)	American Type Culture Collection	ATCC 8482
Clostridium scindens	American Type Culture Collection	ATCC 35704
Clostridium spiroforme	American Type Culture Collection	ATCC 29900
Collinsella aerofaciens	American Type Culture Collection	ATCC 25986
Eubacterium rectale	American Type Culture Collection	ATCC 33656
Parabacteroides distasonis	American Type Culture Collection	ATCC 8503
Prevotella copri	German Collection of Microorganisms and Cell Cultures GmbH	DSM 18205
Experimental models: Cell lines		
Immortalized human keratinocytes (HaCaT)	The German Cancer Research Center (Deutsches Krebsforschungszentrum, DKFZ)	N/A
Human colorectal adenocarcinoma cells (Caco-2)	American Type Culture Collection	HTB-37
Experimental models: Organisms/strains		
Mouse: CD-1	Charles River	18679700-022
Chemicals, peptides, and recombinant proteins		
Luria-Bertani broth	BD	244620
Tryptic soy broth	Sigma	T8907-1KG
Agar	Sigma	05039
MacConkey agar	RPI	M42560-500.0
Brain heart infusion	Remel	R452472
Brain heart infusion broth	Fluka	53286-500G
Phosphate buffer saline	Sigma	P3913-10PAK

(*Continued on next page*)

Continued

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Glucose	Sigma	G5767
1-(N-phenylamino)naphthalene	Sigma	104043
3,3'-dipropylthiadicarbocyanine iodide	Sigma	43608
HEPES	Fisher	BP310-100
Potassium chloride (KCl)	Sigma	P3911
Vitamin K3	Acros Organic	127180250
Hemin	Alfa Aesar	A11165
L-cysteine	Alfa Aesar	A10435
Resazurin	Acros Organic	4189900050
SUPERase-In™ RNase Inhibitor	Invitrogen	AM2694
Quick-RNA Fungal/Bacterial Miniprep Kit	Zymo Research	R2014
RNA Clean-and-Concentrator-5	Zymo Research	R1013
Isoflurane	Covetrus	029405
Deposited data		
List of smORF-encoded candidate peptides with predicted antimicrobial activity; BLASTn results of the predicted SEPs top hits organisms; List of SEPs identified as transcribed and translated based on normalized read coverage of the SEP coding region in RNA-seq and MetaRibo-seq	This study	https://data.mendeley.com/datasets/5dvzpv5f8c/1
RNA-sequencing data of *P. copri* DSM 18205 was deposited in the Sequence Read Archive (SRA)	This study	BioProject: PRJNA1146138
Software and algorithms		
Prodigal (version 2.6.3)	Hyatt et al.[29]	https://github.com/hyattpd/Prodigal
CD-hit	Fu et al.[65]	http://weizhong-lab.ucsd.edu/cdhit_suite
BLASTn	Altschul et al.[66]	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast
AmPEP	Bhadra et al.[67]	https://cbbio.cis.um.edu.mo/software/AmPEP
SmORFinder	Durrant and Bhatt[68]	https://github.com/bhattlab/SmORFinder
ColabFold	Mirdita et al.[69]	https://github.com/sokrypton/ColabFold
Mol*3D Viewer	Sehnal et al.[70]	https://www.rcsb.org/3d-view
BLASTp	Altschul et al.[66]	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast
TrimGalore	TrimGalore	https://github.com/FelixKrueger/TrimGalore
bwa-mem (Burrows-Wheeler Aligner)	Li and Durbin[71]	https://github.com/lh3/bwa
featureCounts	Liao et al.[72]	https://github.com/ShiLab-Bioinformatics/subread
IGV	Integrative Genomics Viewer	https://igv.org/
Prism (version 10.0.1)	GraphPad	https://www.graphpad.com/
BioRender	BioRender	https://www.biorender.com/

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled upon reasonable request by the lead contact, Cesar de la Fuente-Nunez (cfuente@upenn.edu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

- HMPI-II metagenomes from which SEPs were annotated are publicly available from the NIH Human Microbiome Project: https://www.hmpdacc.org/hmasm2/.
- All original code and DOIs are publicly available and listed in the key resources table. Data S1 file is publicly available in Mendeley Data (https://data.mendeley.com/datasets/5dvzpv5f8c/1) and listed in the key resources table.
- Any additional information required to reanalyze the data reported in this paper is available from the lead contact upon request.

EXPERIMENTAL MODEL AND STUDY PARTICIPANT DETAILS

Bacterial strains and growth conditions

Acinetobacter baumannii ATCC 19606, *Escherichia coli* ATCC 11775, *Escherichia coli* AIC221 [*Escherichia coli* MG1655 phnE_2::FRT (control strain for AIC222)] and *Escherichia coli* AIC222 [*Escherichia coli* MG1655 pmrA53 phnE_2::FRT (polymyxin resistant; colistin-resistant strain)], *Klebsiella pneumoniae* ATCC 13883, *Pseudomonas aeruginosa* PAO1, *Pseudomonas aeruginosa* PA14, *Staphylococcus aureus* ATCC 12600, *Staphylococcus aureus* ATCC BAA-1556 (methicillin-resistant strain), *Enterococcus faecalis* ATCC 700802 (vancomycin-resistant strain), *Enterococcus faecium* ATCC 700221 (vancomycin-resistant strain) were grown and plated on Luria-Bertani (LB), Pseudomonas Isolation (*Pseudomonas aeruginosa* strains), and MacConkey (*A. baumannii* strain) agar plates and incubated overnight at 37°C from frozen stocks. After incubation, one isolated colony was transferred to 5 mL of medium (LB), and cultures were incubated overnight (16 h) at 37°C. The following day, inocula were prepared by diluting the overnight cultures 1:100 in 5 mL of the respective media and incubating them at 37°C until bacteria reached logarithmic phase (OD_{600} = 0.3-0.5).

Akkermansia muciniphila ATCC BAA-635, *Bacteroides eggerthi* ATCC 27754, *Bacteriodes fragilis* ATCC 25285, *Bacteroides ovatus* ATCC 8483, *Bacteroides thetaiotaomicron* ATCC 29148, *Bacteroides uniformis* ATCC 8492, *Bacteroides vulgatus* ATCC 8482 (*Phocaeicola vulgatus*), *Clostridium scindens* ATCC 35704, *Clostridium spiroforme* ATCC 29900, *Collinsella aerofaciens* ATCC 25986, *Eubacterium rectale* ATCC 33656, *Parabacteroides distasonis* ATCC 8503, and *Prevotella copri* DSM 18205 were the gut commensal strains used in this study. All the commensal microorganisms were cultured in brain heart infusion (BHI) broth and agar plates enriched with 0.1% (v:v) vitamin K3 (1 mg mL^{-1}), 1% (v:v) hemin (1 mg mL^{-1}, diluted with 10 mL of 1 N sodium hydroxide), and 10% (v:v) $_L$-cysteine (0.05 mg mL^{-1}), from cryopreserved stocks and incubated overnight at 37°C. Resazurin was used as oxygen indicator. After the incubation period, a single isolated colony was transferred to 3 mL of BHI broth and incubated overnight at 37°C. The next day, inocula were prepared by diluting the bacterial overnight cultures 1:100 in 3 mL of BHI broth and incubated at 37°C until reaching the logarithmic phase (OD_{600} = 0.3-0.5).

Eukaryotic cell culture conditions

Immortalized human keratinocytes (HaCaT – obtained from DKFZ)[49] and human colorectal adenocarcinoma cells (Caco-2 HTB-37, ATCC) were cultured in high-glucose Dulbecco's modified Eagle's medium (glumax DMEM – Gibco 11965092) supplemented with 1% v/v antibiotics (penicillin/streptomycin) and 10% v/v fetal bovine serum (FBS). HaCaT and Caco-2 cell lines were grown at 37°C in a humidified atmosphere containing 5% CO_2.

Skin abscess infection mouse model

To assess the effectiveness of the peptides against *A. baumannii* ATCC 19606, the bacteria were cultured in tryptic soy broth (TSB) medium until reaching an OD_{600} of 0.5. Subsequently, the cells were washed twice with sterile PBS (pH 7.4) and suspended to a final concentration of 5×10^6 colony-forming units (CFU) mL^{-1}. For the in vivo experiments, six-week-old female CD-1 mice were anesthetized with isoflurane and subjected to a superficial linear skin abrasion on their backs. A 20 μL aliquot containing the bacterial load suspended in PBS was then inoculated over the scratched area. The peptides, diluted in water at their MIC value, were administered to the infected area one hour after the infection. Two and four days post-infection, the animals were euthanized, and the skin area with the infection was excised and homogenized for 20 min using a bead beater (25 Hz). The homogenates were then 10-fold serially diluted for CFU quantification. A total of three independent experiments were performed (N=3), with three mice per group (n=9). The skin abscess infection mouse model was revised and approved by the University Laboratory Animal Resources (ULAR) from the University of Pennsylvania (Protocol 806763).

Deep thigh infection mouse model

To induce neutropenia in mice, two intraperitoneal doses of cyclophosphamide (150 mg Kg^{-1}) were administered at a 72-hour interval. One day after the last cyclophosphamide dose, the mice were intramuscularly infected in their right thigh with a bacterial load of 5×10^6 CFU mL^{-1} of *A. baumannii* ATCC 19606, which had been cultured in tryptic soy broth (TSB), washed twice with PBS (pH 7.4), and resuspended to the required concentration. Two hours after infection, the mice were treated with peptides suspended in water via intraperitoneal injection. Before each injection, the mice were anesthetized with isoflurane and their respiratory rate and pedal

reflexes were monitored. Subsequently, we closely monitored the progression of the infection and euthanized the mice accordingly. The infected area was removed two and four days post-infection, homogenized for 20 min (25 Hz) using a bead beater, and 10-fold serially diluted for CFU quantification on MacConkey agar plates. A total of three independent were conducted (N=3), each experimental group consisted of 3 mice (n=9). However, during one of the replicates, an unexpected bacterial growth occurred suddenly, leading to a significant increase in bacterial counts by three orders of magnitude (inoculum 5×10^6 CFU mL^{-1} and counts $\sim10^9$ CFU mL^{-1}) within a single day. As a consequence, all three mice in that replicate perished (n=3), which accounts for the variation in the number of mice (n=6) observed on day 4 post infection for each condition. The deep thigh infection mouse model was revised and approved by the University Laboratory Animal Resources (ULAR) from the University of Pennsylvania (Protocol 807055).

METHOD DETAILS

Generation of list of 444,054 smORF family representatives from multiple human associated metagenomes

We started with the list of 444,054 family representatives previously reported in Sberro et al.[23] The computational methodology used to generate that list of representatives is briefly summarized here. Contigs from 1,773 HMPI-II human-associated metagenomes from four major body sites were downloaded. All ORFs $\geq$15 bp were predicted using MetaProdigal.[29] The list was then filtered to only contain small ORFs $\leq$150 bp, resulting in a set of 2,514,099 smORFs. The proteins encoded by these smORFs were clustered into families using as parameters: -n 2 -p 1 -c 0.5 -d 200 -M 50000 -l 5 -s 0.95 –aL 0.95 –g 1 (family members required to have 40-50% homology; shorter sequences required to be $\geq$95% length of the cluster representative; and the alignment was required to cover $\geq$95% of the longer sequence). This resulted in 444,054 clusters, each of which was assigned a 'cluster representative' by CD-Hit.[65] The cluster representative for each of the 444,054 clusters was used in subsequent parts of our analysis. Hereinafter, we use the family ID interchangeably to refer to the family representative.

Antimicrobial peptide prediction

AmPEP[67] was applied (default parameters) on the 444,054 representatives of families. We considered all peptides with AmPEP score greater than 0.5 generating a list of 11,710 family representatives with likelihood of antimicrobial activity.

Application of SmORFinder to predict high confidence peptides

SmORFinder[68] was run on all metagenomes from the Human Microbiome Project (HMPI-II)[73] as previously reported. We filtered the list of 444,054 family representatives for those that were also called by SmORFinder. This resulted in a list of 38,965 family representatives. We then took the intersection of the list of 38,965 representatives called by SmORFinder and the list of 11,710 antimicrobial representatives called by AmPEP. This yielded the list of 323 representatives that were both SmORFinder positive and had AmPEP score $\geq$0.5. This list was used in our subsequent analysis.

Inclusion and exclusion criteria to select peptides for activity testing

Our approach to selecting peptide sequences from the list of 323 was as follows. First, we applied the following exclusion criteria: (1) we removed from consideration all peptides with more than one cysteine residues, these were considered undesirable candidates due to the tendency of these peptides to oxidize and cross-link aggregating in solution; (2) we removed peptides that have a high mean hydrophobicity owing to the difficulty of chemical synthesis and the tendency of these peptides to aggregate. (3) We chose peptides representing the ''known'' sequence space, these were peptides that shared many features of known AMPs and looked to have high confidence of being antimicrobial. (4) We chose peptides representing the negative sequence space, these were peptides that shared similarity to sequences that are either known to not be antimicrobial or not likely to be antimicrobial (*e.g.*, net negative charge). (5) We chose peptides that represented the edge of the ''known'' sequence space. These were peptides that looked promising (*e.g.*, appear to form α-helices, or contain positive charge or poly-lysine residues), but were different from known sequences. (6) We chose peptides that represented the ''unexplored'' sequence space that would typically be looked over by conventional structure-activity relationship design approaches. These were peptides with unconventional sequences, for example ones that have amino acid residues that are not typically found in AMPs, such as polar uncharged residues (Asn, Gln, Met, Ser, Thr).

In all cases, we chose to test only the representative member of the family. For promising candidates, there is the option 'expand' the family to test the other sequences in the cluster. Given that the families were clustered at the 40-50% homology threshold, we expect peptides within the same family to have dramatic differences in activity. In these cases, we could look at individual members of a family for conserved sites of activity. In one case (family 420019), where the cluster representative could not be synthesized, we took the consensus sequence of the alignment of all 47 family members in family 420019.

Sequence similarity score

To calculate the sequence similarity of two given peptide sequences "*i*" and "*j*", we used the Smith-Waterman algorithm[74] that resulted in their alignment score $SW(i,j)$. The sequence similarity score between these two peptide sequences was defined as the normalized alignment score: $\frac{SW(i,j)}{\sqrt{SW(i,i)*SW(j,j)}} \in [0,1]$. In this case, a higher score means a higher sequence similarity between two peptide sequences.

Peptide synthesis

All peptides used in the experiments were purchased from AAPPTec and synthesized by solid-phase peptide synthesis using the Fmoc strategy.

Minimum inhibitory concentration determination

The 78 SEPs underwent broth microdilution assays to evaluate their *in vitro* antimicrobial activity. The determination of the minimum inhibitory concentration (MIC) values was done by utilizing the broth microdilution technique, wherein a starting inoculum of 2×10^6 cells was introduced into LB, in nontreated polystyrene microtiter 96-well plates. Peptides, prepared as aqueous solutions, were added to the plate and two-fold diluted (1 to 128 μmol L^{-1}). The MIC was identified as the lowest concentration of peptide that completely inhibited the visible growth of bacteria after 20 h of incubation at 37°C. The plates were then analysed using a spectrophotometer at 600 nm. The assays were conducted in triplicate to ensure statistical reliability.

Identification of evidence of transcription and translation of SEPs in published MetaRibo-seq data

We searched for peptides belonging to our list of 323 SEPs in a previously published dataset of RNA-seq and associated MetaRibo-seq data.[28] Families with reads per kilobase million (RPKM) values in RNA-seq and MetaRibo-seq data were taken to present evidence of transcription and translation.

RNA-sequencing of *Prevotella copri* DSM 18205

P. copri DSM 18205 was streaked to isolation on a BHI agar plate (supplemented with 0.1% vitamin K3, 1% hemin, and 10% $_L$-cysteine) and incubated for 48 h at 37°C in a Bactron 300 anaerobic chamber (Sheldon Manufacturing Inc., Cornelius, OR). A single colony was picked to inoculate 5 mL of pre-reduced, supplemented BHI media, which was cultured for 48-72 h. This 'starter' population was subsequently sub-cultured into three biological replicates by diluting 1 mL of culture with 4 mL of pre-reduced, supplemented BHI. Each replicate was further incubated anaerobically at 37°C until reaching a predetermined density, at which point 4 mL per replicate was quenched with 500 μL of ice-cold 10% acidic phenol in ethanol, spun down (5000 xg, 10 min, 4°C), and stored at -80°C. This process was repeated three times to collect pellets from replicate populations with optical densities (OD_{600}) of approximately 0.4, 0.6, and 0.9, representing exponential to early-stationary phases of growth. Within a day of collection, frozen pellets were resuspended in 200 μL of a phosphate-buffered saline pre-lysis buffer (containing 0.5 mg mL^{-1} lysozyme and 0.1 U μL^{-1} SUPERase-In™ RNase inhibitor) and incubated at room temperature for 10 min on a benchtop vortexer set to maximum speed. Samples were subsequently processed with the Quick-RNA Fungal/Bacterial Miniprep Kit (Zymo Research) following the manufacturer's instructions, with on-column DNase treatment. The resulting total RNA were cleaned with the RNA Clean-and-Concentrator-5 Kit (Zymo Research) before quantification with Qubit and qualification with Nanodrop and Bioanalyzer assays. Cleaned RNA samples with RIN $\geq$6.0 were subjected to rRNA depletion using Illumina Ribo-Zero Plus followed by stranded library prep with the NEBNext® Ultra II Directional RNA kit and paired-end 150 bp sequencing on a NovaSeq X instrument (performed by Novogene). Reads were trimmed of adapters and low quality bases with Trim Galore (github.com/FelixKrueger/TrimGalore) and aligned to the *P. copri* DSM 18205 reference genome (GCF_020735445.1) with the Burrows Wheeler Aligner (default bwa-mem).[71] Read counts for each gene annotation were generated with featureCounts[72] and normalized. Finally, alignments were visualized with the Integrative Genomics Viewer (IGV) to assess read density across the rpsO locus.

SEP identification in hCom2

119 hCom2 reference genomes, sequenced as part of the original hCom2 publication,[59] were either downloaded from RefSeq as assemblies or assembled with SPAdes[75] (–isolate) from their raw DNA sequencing reads, depending on availability. A custom configuration of Prodigal[29] was run on each assembly to predict all ORFs of at least 15 base pairs in length, and the resulting protein sequences (531,822 in total) were concatenated into a single hCom2 ORF list and converted into a BLAST database (makeblastdb). Short protein-optimized BLAST (blastp-short) was subsequently used to query for significant alignments (E-value $\leq$ 1e-4) between our 323 predicted antimicrobial SEPs and all hCom2 ORFs. For each SEP, the top hit (*i.e.*, lowest E-value) was considered for further analysis, first by filtering out all SEPs with incomplete (i.e., less than 100% amino acid identity across the entire length of the SEP) alignments. The list of complete SEP-hCom2 hits were assessed for in-frame overlap with longer (non-smORF) proteins by searching for each SEP sequence in the Prodigal output of its respective 'hit' organism. Finally, the list of SEPs without overlapping in-frame annotations were manually assessed for out-of-frame overlapping proteins using SnapGene software (www.snapgene.com) and either existing RefSeq annotations or Bakta[76] generated annotations.

Circular dichroism assays

Circular dichroism assays were performed as previously described.[11] Briefly, the experiments were conducted at the University of Pennsylvania's Biological Chemistry Resource CEnter (BCRC) using a J1500 circular dichroism spectropolarimeter (Jasco). The experiments were carried out at 25°C, and the circular dichroism spectra represent an average of three accumulations. Three spectra accumulations were obtained using a quartz cuvette with an optical path length of 1.0 mm, covering a wavelength range from 260 to 190 nm at a rate of 50 nm min^{-1} and a bandwidth of 0.5 nm. The concentration of all peptides tested was 50 μmol L^{-1}, and the measurements were performed in a mixture of trifluoroethanol (TFE) and water in a 3:2 ratio. Respective baselines were recorded before

taking the measurements, and a Fourier transform filter was applied to minimize background effects. Secondary structure fraction values were calculated using the single spectra analysis tool on the server BeStSel.[42]

Synergy assays

The selection of *A. baumannii* ATCC 19606 for the synergy assays was based on its significance as a pathogen that exhibits intrinsic resistance to antimicrobial agents and its capacity to infect various sites including the urinary tract, gastrointestinal tissue, and skin and soft tissues.[55] Following determination of the minimum inhibitory concentration (MIC) for each peptide, the most efficacious SEPs against *A. baumannii* were subjected to orthogonal dilution and concentration range from 2×MIC to 0.03×MIC. The plates were then incubated at 37°C for 24 h. All assays were conducted in triplicate to ensure reliability of results.

Cytoplasmic membrane depolarization assays

The determination of the peptides' ability to depolarize the cytoplasmic membrane was carried out by measuring the fluorescence of the membrane potential-sensitive dye, 3,3'-dipropylthiadicarbocyanine iodide ($DiSC_3$-5). For this experimental protocol, we cultured *A. baumannii* ATCC 19606, *E. coli* ATCC 11775, *E. coli* AIC221, *E. coli* AIC222, *K. pneumoniae* ATCC 13883, *P. aeruginosa* PA14, *S. aureus* ATCC 12600, methicillin-resistant *S. aureus* ATCC BAA-1556, vancomycin-resistant *E. faecalis* ATCC 700802, and vancomycin-resistant *E. faecium* ATCC 700221 with agitation at 37°C until OD_{600} = 0.5. Subsequently, the cells were centrifuged and washed twice with washing buffer (20 mmol L^{-1} glucose, 5 mmol L^{-1} HEPES, pH 7.2) and diluted 10-fold in the same buffer containing 0.1 mol L^{-1} KCl. Next, the cells (100 μL) were incubated for 15 min with $DiSC_3$-5 (20 nmol L^{-1}) until the fluorescence stabilization. We then incubated the peptides (100 μL solution at MIC values) and the changes in fluorescence emission intensity of $DiSC_3$-5 (λ_{ex} = 622 nm, λ_{em} = 670 nm) were used to track depolarization of the cytoplasmic membrane for 60 min. The relative fluorescence was calculated using a non-linear fit and the positive control (buffer + bacteria + fluorescent dye + polymyxin B) served as baseline for comparison. The following equation was applied to determines the % difference between the baseline and the sample:

$$Percentage\ difference = \frac{100 * (fluorescence_{sample} - fluorescence_{polymyxin\ B})}{fluorescence_{polymyxin\ B}}$$

Outer membrane permeabilization assays

The membrane permeability of SEPs was assessed using the 1-(N-phenylamino)naphthalene (NPN) uptake assay. *A. baumannii* ATCC 19606, *E. coli* ATCC 11775, *E. coli* AIC221, *E. coli* AIC222, *K. pneumoniae* ATCC 13883, and *P. aeruginosa* PA14 were cultured until reaching an OD_{600} of 0.4. The cells were then centrifuged (10,000 rpm at 4°C for 10 min), washed, and resuspended in a buffer (5 mmol L^{-1} HEPES, 5 mmol L^{-1} glucose, pH 7.4). Four μL of NPN solution (0.5 mmol L^{-1}) was added to the bacterial solution in a white 96-well plate (total volume of 100 μL). The baseline fluorescence, i.e., fluorescence before the addition of peptides, was measured at λ_{ex} = 350 nm and λ_{em} = 420 nm. Membrane permeability was monitored after the addition of peptides at their MIC (100 μL) until no further increase was observed (for 45 min). The relative fluorescence was calculated using a non-linear fit and the positive control (buffer + bacteria + fluorescent dye + polymyxin B) was used as the baseline. The following equation was used to determine the % difference between the baseline and the sample:

$$Percentage\ difference = \frac{100 * (fluorescence_{sample} - fluorescence_{polymyxin\ B})}{fluorescence_{polymyxin\ B}}$$

Cytotoxicity assays

Cells were seeded into 96-well plates at the density of 5×10^3 cells per well for 24 h. After this period, we treated them with increasing concentrations of peptide (4-64 μmol L^{-1}). Next, MTT reagent at 0.5 mg mL^{-1} in DMEM medium without phenol red was used to replace cell culture supernatant (100 μL per well). The samples were incubated for 4 h at 37°C to obtain the insoluble formazan salts. Then, the resulting salts were solubilized in 0.04 mol L^{-1} hydrochloric acid (HCl) in anhydrous isopropanol and quantified using a spectrophotometer at 570 nm.

QUANTIFICATION AND STATISTICAL ANALYSIS

Reproducibility of the experimental assays

All assays were performed in three independent biological replicates. Cytotoxic activity values were determined through non-linear regression analysis, utilizing a peptide gradient of concentrations. These values represent the concentrations required to kill 50% of the cells in the experiment. For the cytotoxic activity assays, two technical replicates were conducted within each of the three biological replicates. In the skin abscess and deep thigh infection mouse models, we employed three mice per group in three independent replicates, adhering to established protocols approved by the University Laboratory of Animal Resources (ULAR) at the University of Pennsylvania.

Statistical tests

In the mouse experiments, the statistical significance was determined using one-way ANOVA followed by Dunnett's test. All the p values are shown for each of the groups, all groups were compared to the untreated control group.

Statistical analysis

All calculation and statistical analyses of the experimental data were conducted using GraphPad Prism v.10.0.1. Statistical significance between different groups was calculated using the tests indicated in each figure legend. No statistical methods were used to predetermine sample size.

CellPress
OPEN ACCESS

Supplemental figures

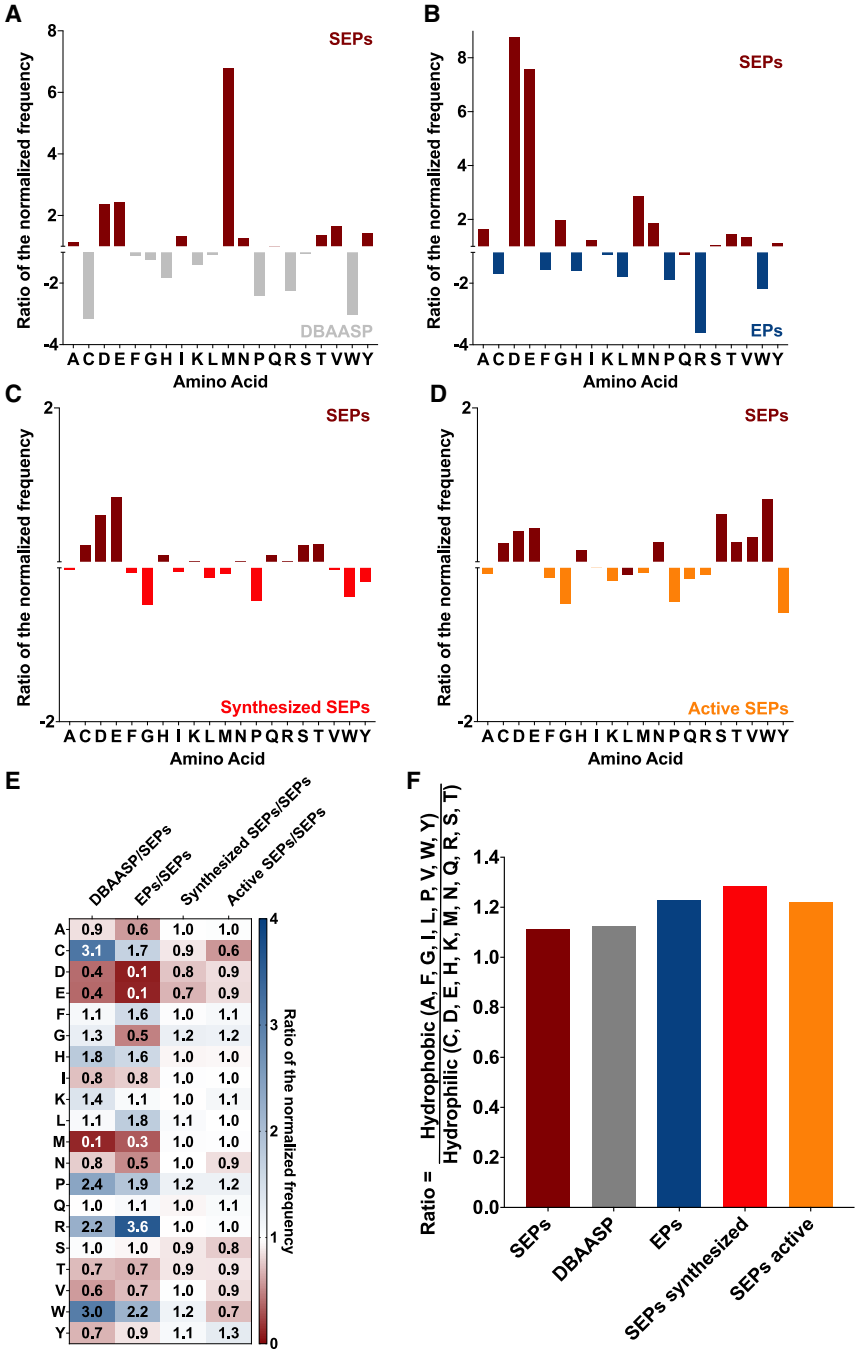
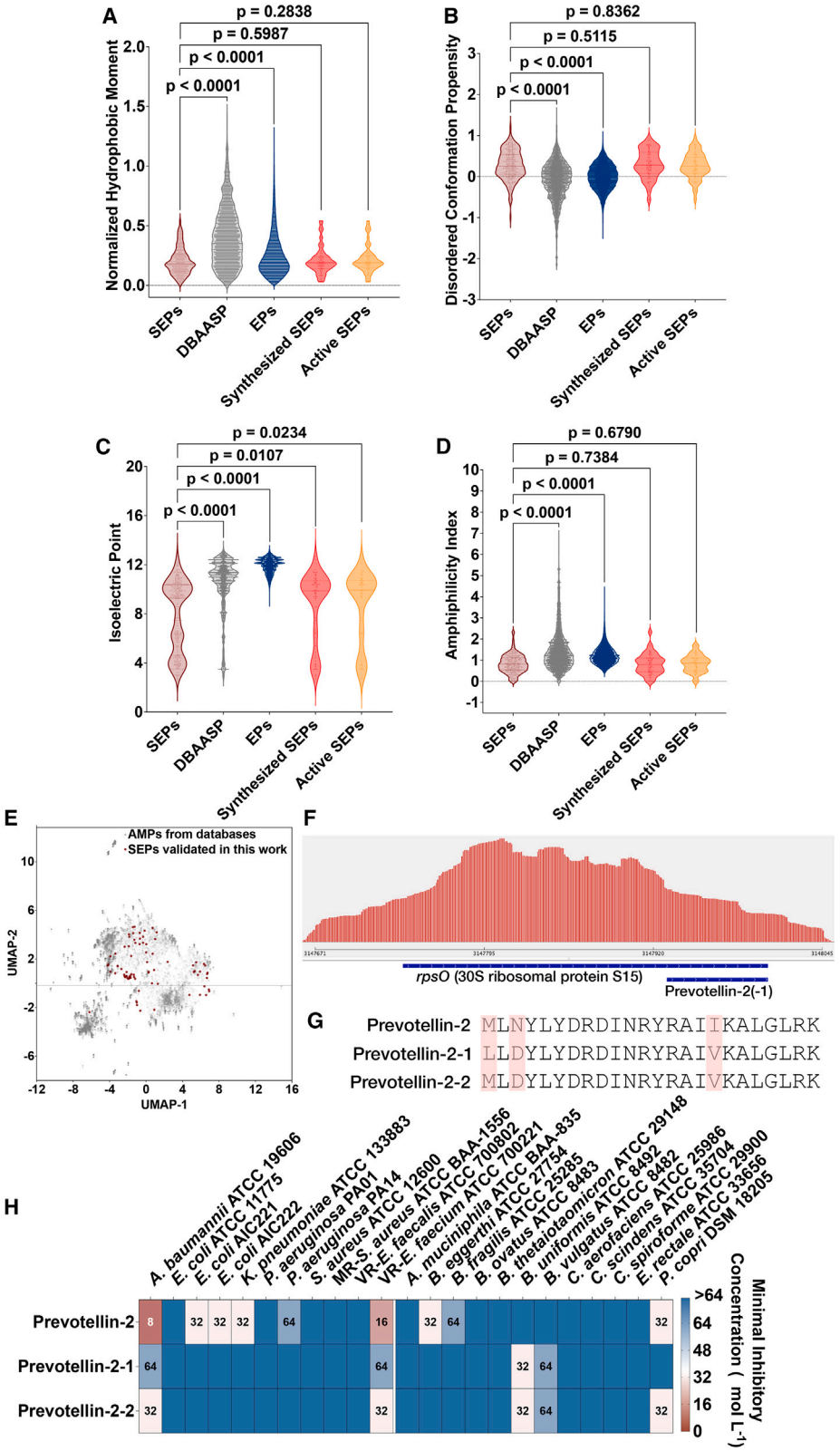


Figure S1. Frequency distribution by amino acid and amino acid type of SEPs compared with known AMPs and EPs from the human proteome, related to Figure 2

(A–D) Ratio of the normalized frequency of SEPs compared with (A) AMPs from DBAASP, (B) EPs from the human proteome, (C) synthesized SEPs, and (D) SEPs experimentally validated as antimicrobially active.

(legend on next page)

(E) Normalized frequency of amino acid type showing that the SEPs present more negatively charged and polar uncharged residues than known AMPs and EPs, whereas AMPs and EPs have more positively charged and aromatic residues than EPs. These results also show that SEPs that were synthesized and SEPs validated as active are very similar in amino acid type content, except for the lower content of cysteine and tryptophan in active SEPs.
(F) Ratio between hydrophilic and hydrophobic amino acid residues for each one of the different classes of peptides analyzed in this work (SEPs, synthesized SEPs, active SEPs, AMPs form DBAASP, and EPs from the human proteome). SEPs are more hydrophobic than the other two classes. All 323 SEP and 43,000 EP candidates were used for the frequency calculation. Synthesized SEPs comprised of a subgroup of 78 molecules, and there was a total of 55 active SEPs among the 78 synthesized ones.



(legend on next page)

Figure S2. Physicochemical features of SEPs compared with antimicrobial peptides and encrypted peptides from the human proteome, sequence space exploration using a similarity matrix, and investigation of prevotellin-2, related to Figures 2 and 3

(A–D) The following physicochemical features were estimated using the Database of Antimicrobial Activity and Structure of Peptides (DBAASP) server[32]: (A) normalized hydrophobic moment, (B) disordered conformation propensity, (C) isoelectric point, and (D) amphiphilicity index. Those physicochemical features summed to net charge (Figure 2B) and normalized hydrophobicity (Figure 2C) are the most relevant easy-to-extract properties that influence the antimicrobial activity and toxicity of peptides with antimicrobial properties.

(E) The graph represents a bidimensional sequence space visualization of peptide sequences found in the databases DBAASP, APD3, and DRAMP 3.0 and validated SEPs. We used sequence alignment to generate a sequence similarity matrix for all peptide sequences in the databases, and the 78 SEPs synthesized and experimentally validated in this work. Each row in the similarity matrix corresponds to a feature representation of a peptide in terms of amino acid residues. Uniform manifold approximation and projection (UMAP) was used to reduce the feature representation to two dimensions for visualization purposes.

(F) Representative RNA-seq read density plot across the rpsO and prevotellin-2 locus. Representative data are from a single total RNA sample from type strain P. copri DSM 18205 grown to early-stationary phase (OD_{600} = 0.9).

(G) Amino acid sequences of metagenome-encoded prevotellin-2 and its type-strain (DSM 18205) variants synthesized for this study (differing amino acids in red). prevotellin-2-1 is the exact amino acid sequence in DSM 18205, while prevotellin-2-2 switches the first leucine (L) for the alternative start amino acid methionine (M).

(H) Antimicrobial activity of prevotellin-2 and its two variants.

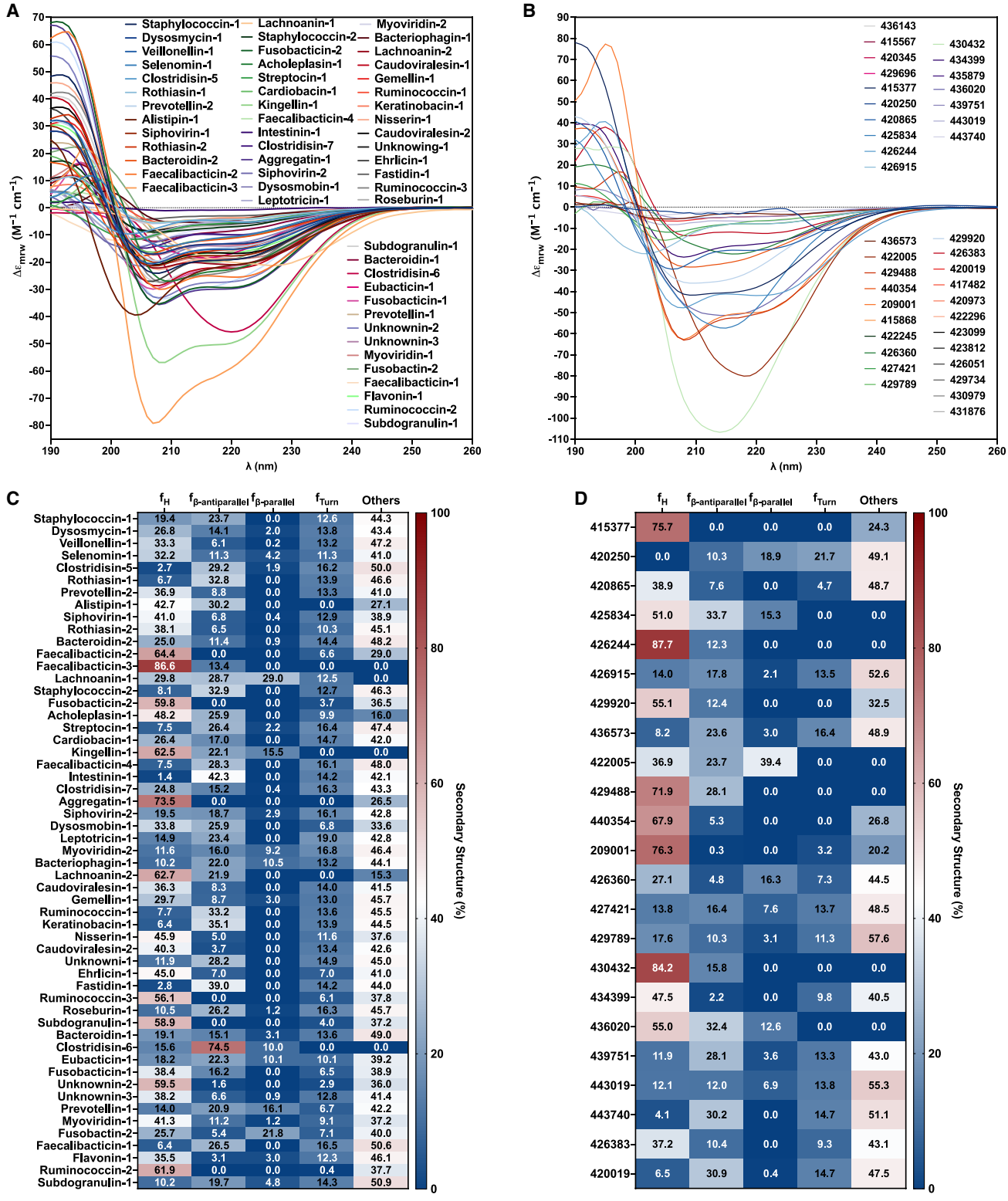


Figure S3. Experimental determination of the secondary structure of active and inactive SEPs, related to Figure 3

Circular dichroism spectra of all (A) active and (B) inactive SEPs in TFE/water (3:2, v:v) synthesized and tested in this work. Secondary structure fractions for the (C) active and (D) inactive SEPs were calculated using the BeStSel server.

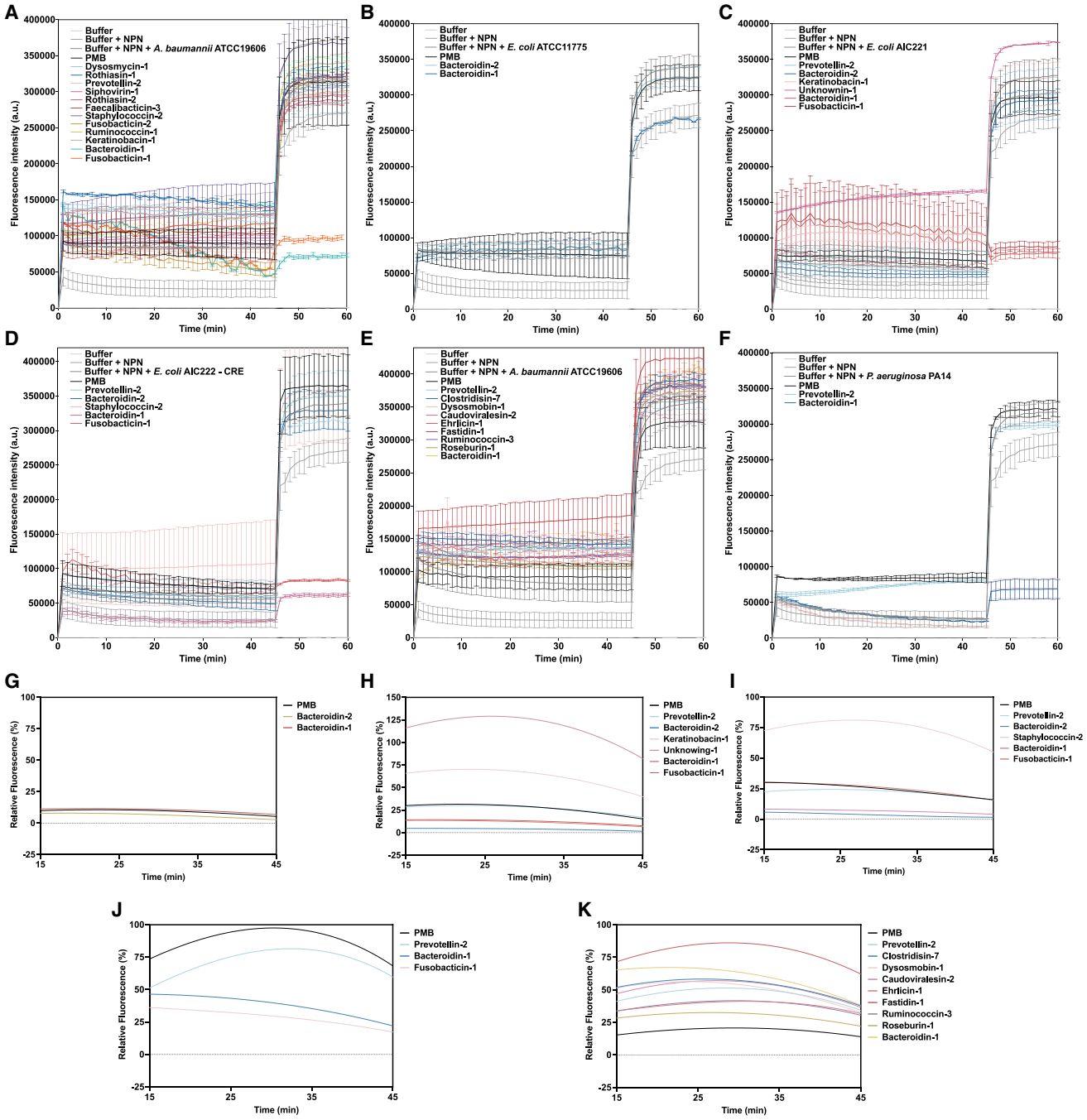
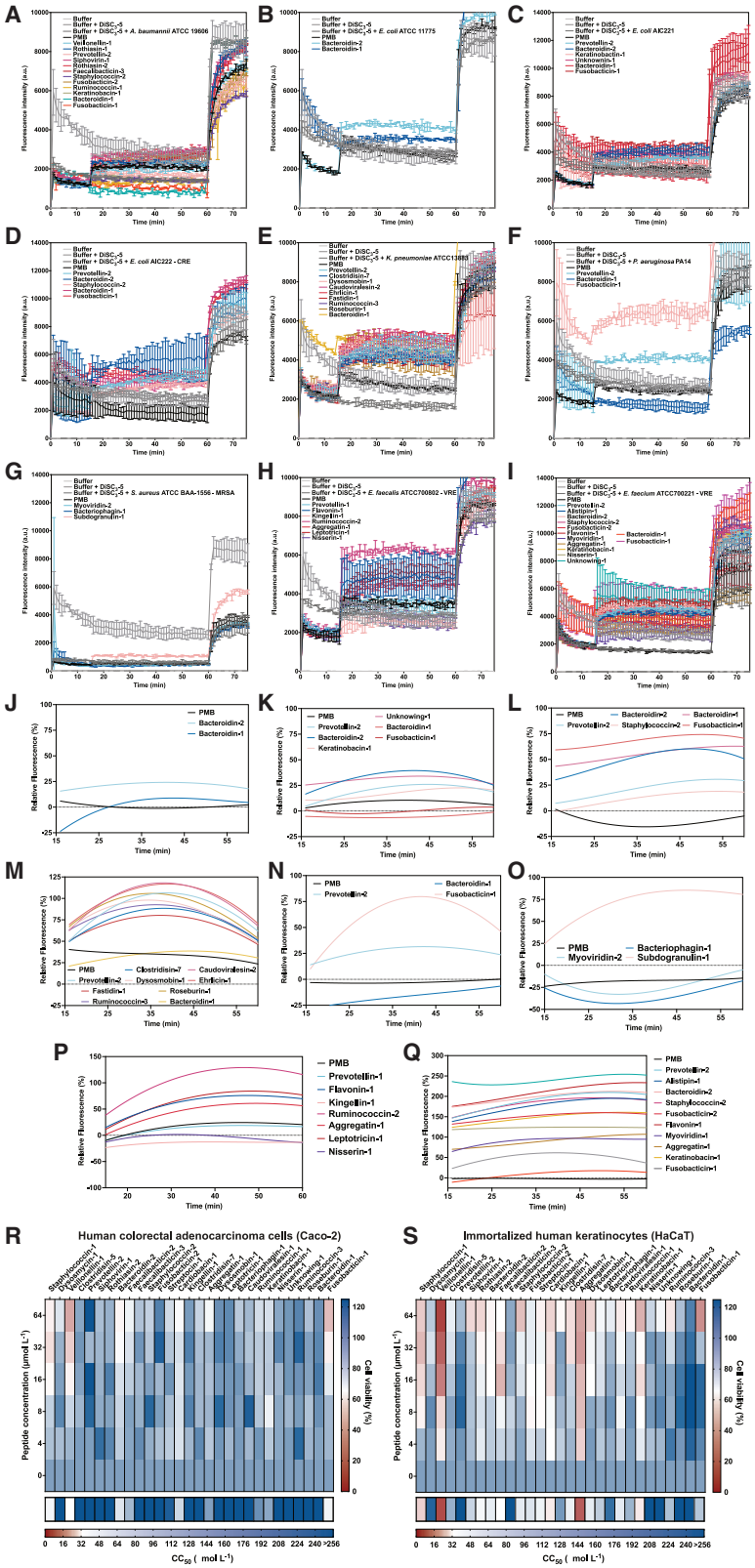


Figure S4. Outer membrane permeabilization caused by SEPs on the membranes of gram-negative pathogenic strains, related to Figure 4

Permeabilization of the outer membrane using the probe 1-(N-phenylamino)naphthalene (NPN) on all the gram-negative strains that were targeted by the SEPs: (A) *A. baumannii* ATCC 19606, (B) *E. coli* ATCC 11775, (C) *E. coli* AIC221, (D) *E. coli* AIC222, (E) *K. pneumoniae* ATCC 13883, and (F) *P. aeruginosa* PA14. Polymyxin B was used as positive control and buffer with NPN, and bacteria were used as baseline for the calculation of the relative fluorescence values (G–K), in experiments against (G) *E. coli* ATCC 11775, (H) *E. coli* AIC221, (I) *E. coli* AIC222, (J) *K. pneumoniae* ATCC 13883, and (K) *P. aeruginosa* PA14. Data in (A–F) are represented as mean and ±SD.



(legend on next page)

Figure S5. Cytoplasmic membrane depolarization triggered by smORF-encoded peptides from on *A. baumannii* and *P. aeruginosa* cell membranes and cytotoxic effects on Caco-2 and HaCaT, related to Figure 4

Depolarization assays with the hydrophobic probe 3,3′-dipropylthiadicarbocyanine iodide ($DiSC_3$-5) on all pathogenic strains targeted by the SEPs: (A) *A. baumannii* ATCC 19606, (B) *E. coli* ATCC 11775, (C) *E. coli* AIC221, (D) *E. coli* AIC222, (E) *K. pneumoniae* ATCC 13883, (F) *P. aeruginosa* PA14, (G) methicillin-resistant *S. aureus* ATCC BAA-1556, (H) vancomycin-resistant *E. faecalis* ATCC 700802, and (I) vancomycin-resistant *E. faecium* ATCC 700221. Polymyxin B was used as positive control and buffer with $DiSC_3$-5, and bacteria were used as baseline for the calculation of the relative fluorescence values (J–Q) in experiments against (J) *E. coli* ATCC 11775, (K) *E. coli* AIC221, (L) *E. coli* AIC222, (M) *K. pneumoniae* ATCC 13883, (N) *P. aeruginosa* PA14, (O) methicillin-resistant *S. aureus* ATCC BAA-1556, (P) vancomycin-resistant *E. faecalis* ATCC 700802, and (Q) vancomycin-resistant *E. faecium* ATCC 700221. Heat maps showing cell viability after 24 h of peptide treatment at concentrations ranging from 4 to 64 μmol L^{-1} against (R) Caco-2 and (S) HaCaT cells. Toxic (in red) and non-toxic (in blue) concentrations of the peptides are a mean of three independent replicates. The row below each of the graphs shows a summary of the predicted CC_{50} concentrations (μmol L^{-1}) of each peptide, i.e., peptide concentrations responsible for 50% of cell death. CC_{50} values have been predicted by interpolating the dose-response with a non-linear regression curve. Data in (A–I) are represented as mean and ±SD.