# Discovery and optimization of antimicrobial peptides from extreme environments on global scale

Zixin Kang[1, #], Haohong Zhang[1, #], Qize Zhou[1, 2, #], Jiayi Liu[1, 2], Kouyi Zhou[1], Peng Chen[1, 2, *], Bi-Feng Liu[1, 2, *], Kang Ning[1, *]

[1] *Department of Bioinformatics and Systems Biology, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, Hubei, China*

[2] *Department of Biomedical Engineering, College of Life Science and Technology, Huazhong University of Science and Technology, Wuhan 430074, China*

[#] These authors contributed equally to this work

[*] Correspondence author

E-mail: gwchenpeng@mail.hust.edu.cn (Peng C.) bfliu@mail.hust.edu.cn (Bi-Feng L.) ningkang@hust.edu.cn (Ning K.)

## Abstract

Novel antibiotics to combat global antimicrobial resistance (AMR) in human and animal pathogens are urgently required [1]. Antimicrobial peptides (AMPs) are a class of small molecules inhibiting growth of various microorganism, including both Gram-negative and Gram-positive bacteria, fungi and viruses [2]. These peptides are valued for their broad-spectrum antimicrobial activity, achieved through mechanisms such as bacterial membrane disruption or interference with intracellular processes rather than targeting specific proteins, which makes lower propensity to induce resistance. While AMPs have been extensively identified and verified from animal proteomes, reference microbial genomes and host environments, those from extreme environments remain unexplored.

Extreme environments, such as deep-sea hydrothermal vents, glaciers, the polar regions, plateau or hot springs, are widely distributed globally and exhibit steep environmental gradients, harsh physiochemical conditions and limited nutrient availability [3-7]. Microbes in such niches evolve unique membrane modification,

specialized metabolic pathway and other strategies to cope with the extreme stresses [3]. Natural geographic isolation and competitive pressures driven by extreme environments make extremophiles an ideal reservoir for mining novel antibiotics. Archaea, which possess vast antimicrobial potential through the production of unique compounds such as archaeasins [8] that exhibit activity against a range of drug-resistant bacteria, often dominate these habitats [4, 9]. Although recent studies have employed deep learning models to mine archaeal proteome [8], the antimicrobial potential of uncultured archaea and other extremophiles on a global scale remains largely untapped.

Many computational approaches have been developed for in-silico screening of AMPs. Traditional approaches encode sequences with physicochemical and biochemical properties to encode sequences [10-13], whereas some recent methods leverage protein language models (pLMs) to generate diverse representations [14-16]. These encoding strategies are typically combined with deep learning models, such as Long Short-Term Memory (LSTM) networks [17] or attention-based models [14, 18, 19], to predict peptide antimicrobial activities. Beyond identification, generative models have been developed to design novel AMPs *de novo* [20-23], and optimization can be achieved through genetic algorithms [24] or reinforcement learning (RL) [19, 22]. However, existing approaches often overlook structural information of peptides and frequently require training complex decision models for virtual evolutions.

To overcome the bottleneck of traditional data mining sources and limitations in peptide screening and optimization, here we introduced structure-aware extremophile genome mining for antimicrobial peptides (SEGMA), a deep-learning based framework used to systematically mine all AMPs from extreme environments on the global scale, herein referred to as extremocin. By leveraging computational pipeline, we identified 3,298 extremocins from 60,461 extremophile metagenome-assembled genomes (MAGs).

## Results

### A smORF catalog of global extremophiles

The global extremophile smORF catalog was derived from 60,461 metagenome-assembled genomes (MAGs) spanning multiple extreme environments worldwide, including polar regions, the Tibet Plateau, glacier-fed streams, deep-sea habitats, and other unique ecosystems including Mariana Trench and Tengchong hot springs (**Fig. 1**). Among 1,406 metagenome samples, over 400 samples were obtained from Tibet Plateau ecosystems, encompassing lake, wetland and glacier in two regions, Tibet and Qilian. The Tibet Plateau developed unique microbial community patterns in response to its extreme conditions, including high elevation, low oxygen concentrations, intense UV radiation, drastic temperature fluctuations, and prolonged

geological isolation [25, 26]. The metagenomic samples span an altitudinal range of 15 kilometers and a temperature gradient of 120 °C (**Fig. 1b**). A total of 3,784,363 smORFs were predicted by modified Prodigal (**Fig. 1c ~ d**). The smORF was defined as nucleic acid sequence encoding small protein up to 50 amino acids. We clustered these smORFs at 50% sequence identity and 95% alignment coverage, resulting in 2,140,838 clusters, which are referred to as families.

To investigate the presence of smORFs in our catalog compared to those in reference microbial genomes from the Genome Taxonomy Database (GTDB), we calculated the number of smORFs per megabase pair (Mbp), termed smORF density (**Supplementary Fig. 1**). Overall, extremophile bacteria exhibit higher smORF density than their counterparts in GTDB ($P_{Mann} < 0.0001$). The bacterial phyla showing the highest density are *Desantisbacteria* (63.96 smORFs per Mbp) and *Edwardsbacteria* (58.27 smORFs per Mbp). The archaeal phyla, *Nanohaloarchaeota* (60.27 smORFs per Mbp) and *Undinarchaeota* (56.88 smORFs per Mbp), exhibit the highest density. Different from the existing study and GTDB results, which suggest that archaea harbor more smORFs, extremophile archaea exhibit a density distribution similar to bacteria ($P_{Mann} = 0.0026$). Besides, we matched small proteins encoded by extremophile smORFs to the Uniref90 database [27] and previously published small protein family datasets, including smprot [28] and GMSC [9]. Only 18.5% of smORFs in our catalog are homologous to these previously reported small proteins (**Fig. 1e**).

## Overview of SEGMA

Building on the global extremophile catalog, we propose the SEGMA framework (Structure-aware Extremophile metaGenome Mining for Antimicrobial peptides). The main idea is to detect and optimize extremocins from the catalog with peptide structure constraints to achieve a high success rate and broad-spectrum antimicrobial activity (**Fig. 1**). SEGMA consists of five modules: APEX 1.1 [8], ESMFold [29], Foldseek [30], SEGMA-Prospect and SEGMA-Search. Candidate extremocins were first prioritized with APEX 1.1, and were further refined with SEGMA-Prospect. It employed a structure-aware protein language model, SaProt-650M, to integrate peptide sequence and structural features, with structures represented as 3Di sequences generated by Foldseek coupled with ESMFold (see **Methods**). Due to the high novelty and diversity of these peptides compared to the known sequence universe (**Fig. 1e**), SaProt-650M was fine-tuned on our catalog and peptides from Uniprot, using Low Rank Adaptation (LoRA), which enables the model to capture biome-specific evolutionary information while avoiding catastrophic forgetting of pretrained knowledge [31-33]. Then, a Gradient Boosting tree model (GBM) was trained on peptide embedding to predict AMP probability.

SEGMA-search frames peptide optimization as a beam search process, where candidate mutations are guided by the language model, and mutation actions are

selected through APEX 1.1 and manually defined rules (**Fig. 1f**). To retain biome-specific peptide space while capturing antibiotic features, SaProt-650M was further fine-tuned on DBAASP, a curated database of experimentally validated AMPs. To comprehensively explore the peptide fitness landscape, residues at each position of peptides were iteratively masked and predicted while preserving complete structure information. The fine-tuned model generated a mutation score map, and residues with the top-30 scores were selected as new candidates. These were subsequently evaluated with APEX 1.1 to estimate MIC values. The top-10 candidates were retained for iterative rounds of optimization.

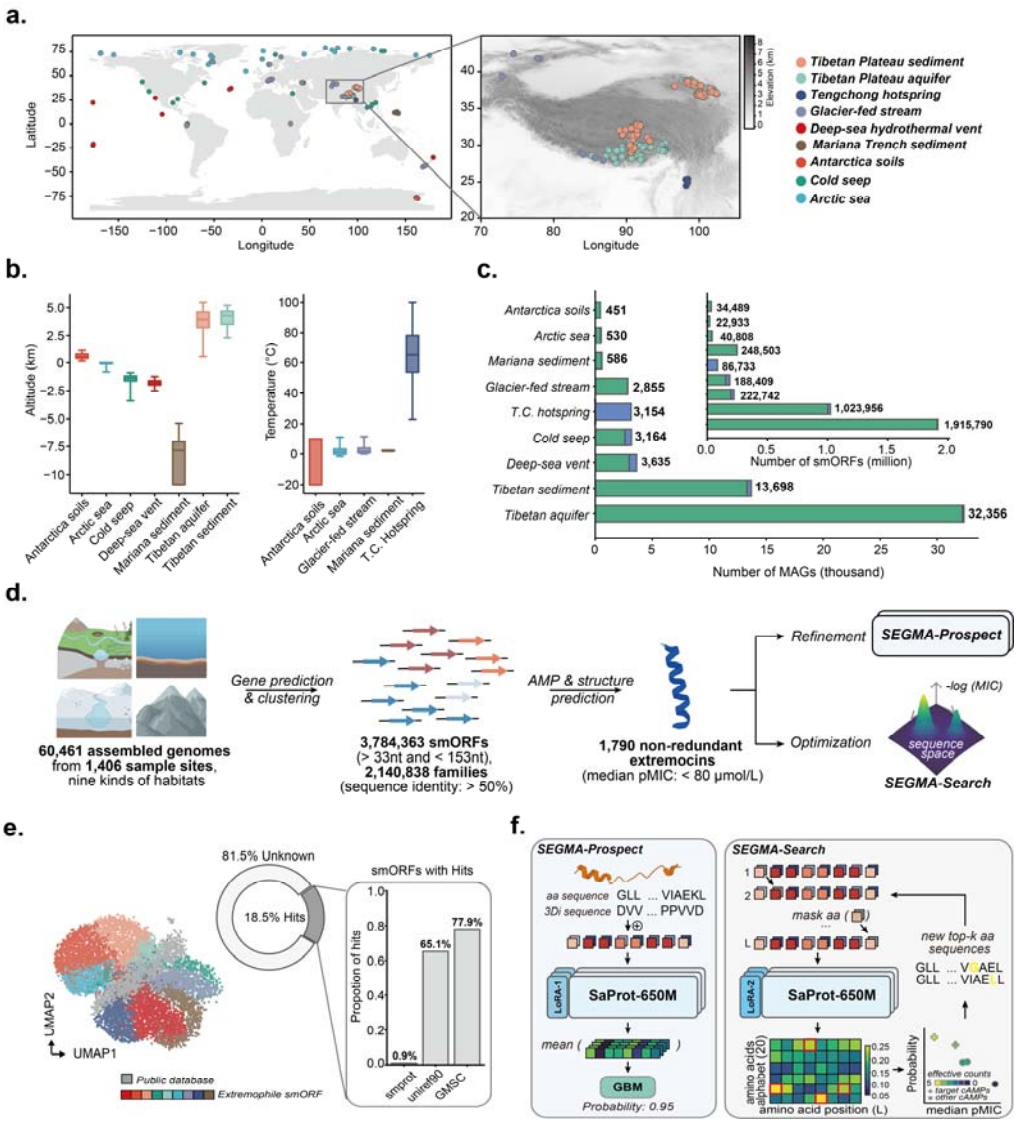



**Figure 1. The global extremophile catalog and the workflow of SEGMA. a.** The geographic distribution of 1,406 metagenome samples from nine kinds of habitats: Tibetan Plateau aquifer, Tibetan Plateau sediment, Arctic Sea, Antarctic soil, Global glacier-fed streams, Tengchong hot springs, Global deep cold seep, Global deep-sea hydrothermal vent and Mariana Trench sediments. **b.** The altitude and temperature

distribution of sample sites. The altitude of different sampling points spans 15 kilometers and the temperature spans 120 degree. **c.** Number of metagenome-assembled genomes (MAGs) across nine habitats and **d.** To build SEGMA, (i) 60,461 metagenomes were collected. (ii) A total of 3,784,363 smORFs (33 bp ~153 bp) were predicted and clustered into 2,140,838 families at 50% identity cutoffs. (iii) Antimicrobial activity was predicted using APEX 1.1, and peptides showing broad-spectrum activity with a median MIC ≤ 80 μmol/L were retained as extremocins, yielding 1,790 non-redundant candidates. **e.** UMAP visualization of smORFs from nine habitats and a public catalog, GMSC, using DNA k-mer encoding (k = 7). Each point represents a sequence. Only 18.5% smORFs have homologous sequences in other sequence catalogs. **f.** Workflow of SEGMA-Prospect: (iv) To incorporate structure information, ESMFold was used to predict peptide structure, which were then tokenized into 3Di sequences with Foldseek. Combined amino acid and 3Di sequences were embedded using a fine-tuned SaProt-650M model. A Gradient Boosting Model (GBM) was then applied to estimate AMP probability. Workflow of SEGMA-Search: (v) To explore the entire fitness landscape, residues at each position were systematically masked and predicted while preserving complete structure information. Coupled with beam search strategy, interactive single-mutation enabled extremocins to evolve from narrow-spectrum to broad-spectrum antimicrobial activity.

## Identification of extremocins from metagenomes

By mining 60,461 MAGs from 1,406 extremophilic habitats spanning multiple geographical scales, we identified 3,298 candidate extremocins (median MIC ≤ 80 μmol/L predicted by APEX1.1), comprising 1,790 non-redundant sequences assigned to 1,390 gene families, few of which match existing AMP databases (AMPSphere, DBAASP) (**Fig. 2**). To assess the taxonomic composition, we annotated the MAGs containing extremocins using GTDB-Tk, assigning them to 576 bacterial genera and 139 archaeal genera (**Fig. 2a**). The bacterial phyla contributing the most extremocins were Bacteroidota (697), Proteobacteria (371) and Patescibacteria (154), all of which were dominant in most habitats, including the Tibet Plateau, cold seeps and glacier-fed steams. The archaeal phyla contributing the most were Thermoproteota (657), Thermoplasmatota (170) and Halobacteriota (106), which were mainly enriched in Tengchong hot springs and deep-sea hydrothermal vents characterized by high temperature (**Fig. 2b**).

To quantify phylum-specific antimicrobial potential, we calculated extremocin density as the number of extremocins per gigabase pair of assembled sequence. Notably, archaeal phyla harbored higher densities than bacterial phyla ($P_{Mann} < 0.0001$), rather than showing similar distributions of smORF densities (**Fig. 2c, Supplementary Fig.2a**). Among bacterial phyla, Patescibacteria exhibited the highest density, which also referred to as Candidate Phyla Radiation (CPR) group featured with

small genomes and absence of biosynthetic pathways [34-36]. To further examine the distribution of sequence families across the ten dominant phyla ($n \geq 64$), we calculated phylum-specific extremocin diversity and specificity (see **Methods**). Bacterial phyla exhibit higher diversity and specificity than archaeal phyla.

## Expansion of antibiotic space by extremocins

To investigate differences between SEGMA-specific extremocins (EPs; extremocins present only in our SEGMA's catalog) and reference-extremophile extremocins (REPs; cAMPs present in both the extremophile and reference genome catalogs), we compared all extremocins with AMPs predicted from GTDB genomes, resulting in 2,765 EPs and 533 REPs.

To determine whether EPs constitute novel categories of antimicrobial peptides, we compared sequence features across EPs, REPs and conventional AMPs from DBAASP. Using UMAP (Uniform Manifold Approximation and Projection) for dimensionality reduction on the sequence similarity score matrix, we observed that DBAASP AMPs primarily cluster in the central region of the sequence space, whereas EPs and REPs encompass a broader spatial distribution, with EPs forming diverse clusters uncovered by REPs or DBAASP entries (**Fig. 2d**). To further investigate the sources of these outlier clusters, we colored the UMAP points according to the habitats or the database. Metagenomes from Tengchong hot spring and Tibet Plateau contributed most, which also harbored higher proportion of EPs and dominated diverse cAMP families, underscoring their vast potential for antibiotic discovery (**Supplementary Fig. 3a, d, e**).

Furthermore, the amino acid composition (**Fig. 2e**) and physicochemical property (**Fig. 2f, Supplementary Fig. 4**) analysis revealed distinct differences between extremocins (EPs & REPs) and DBAASP AMPs. Extremocins were significantly enriched in cationic amino acids, including lysine and arginine, and exhibited substantially lower proportions of glycine, suggesting a charge distribution balance consistent with prior research about archaeasins [8] (**Fig. 2e**). Despite this balance, both EPs and REPs possessed higher net charges than DBAASP AMPs, indicating enhanced electrostatic interactions with negatively charged bacterial membranes (**Fig. 2f**). Notably, we observed that the proportion of arginine in EPs was lower than in REPs, whereas the proportion of glutamate, the precursor in arginine synthesis [37, 38], was higher than in REPs. We inferred that the difference may stem from limited nitrogen availability in extreme environments, which restricts the transfer of the acetyl group from N-α-acetylornithine to glutamate. Moreover, EPs were enriched in phenylalanine and leucine, resulting in greater hydrophobicity and lower boman index than REPs, although both were less hydrophobic than DBAASP AMPs (**Fig. 2f, Supplementary Fig. 4**). These findings indicate that extremocins preferentially bind to membranes via electrostatic interaction, highlighting their potential as cationic antibiotics targeting diverse drug-resistant Gram-negative pathogens [39, 40].

Given the distinct sequence features of EPs and REPs, we investigated their function proportions using eggNOG-mapper [41]. Only 38.17% of extremocins are annotated with COG category, 878 (31.75%) of which are EPs and 381 (71.48%) are REPs. Most EPs cannot be assigned to the specific categories. The COG classes that belong to Information storage and processing predominated in both EPs and REPs, consistent with prior studies [9, 18, 42]. Furthermore, EPs exhibit a greater COG category diversity than REPs, encompassing cellular processing and signaling, as well as metabolism and transport. This suggests that genes spanning diverse functional classes in extremophilic microorganisms possess potential antimicrobial activities.

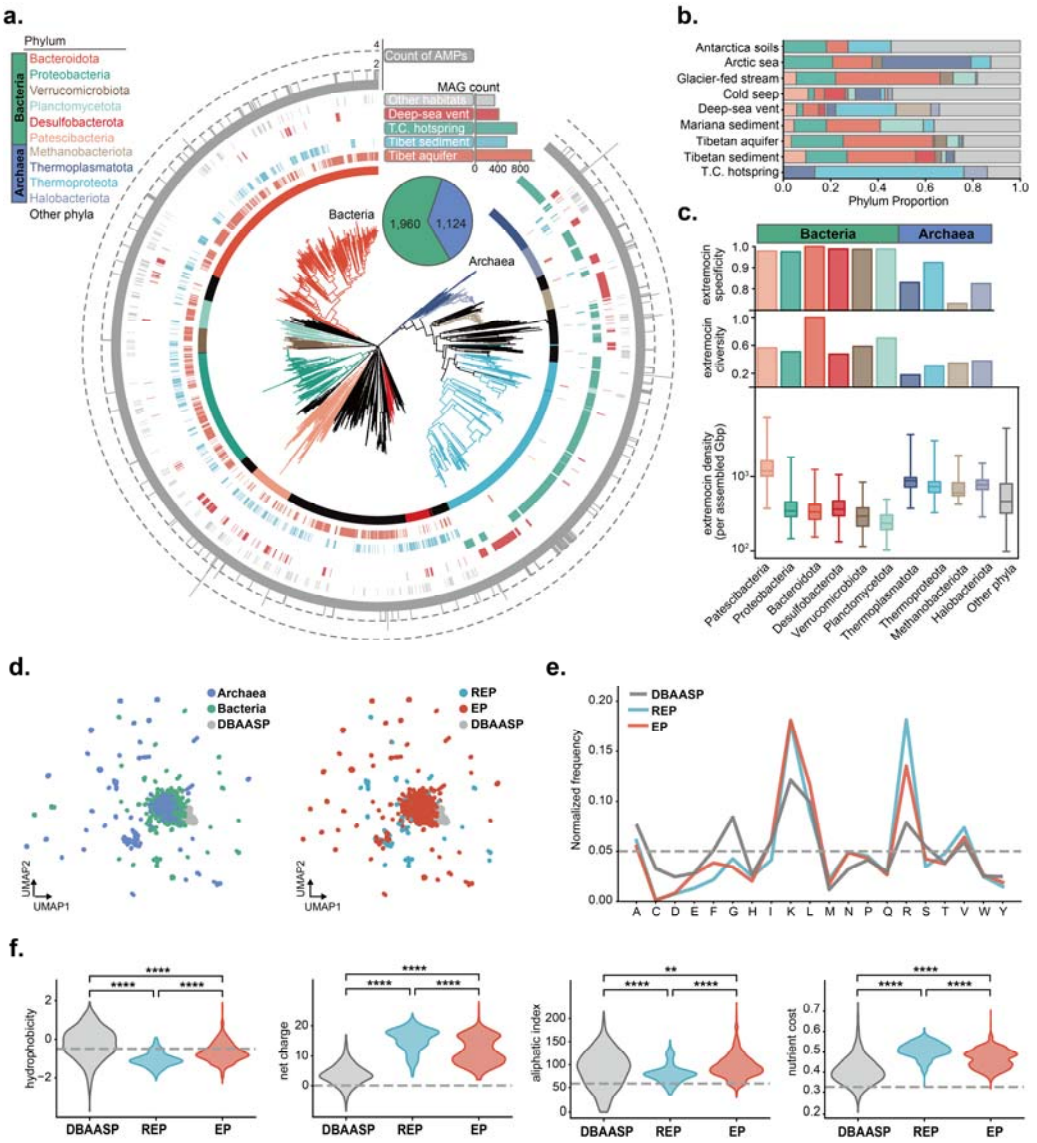



**Figure 2. Distributions, sequence features and physiochemical properties of extremocins. a.** The phylogenetic tree of extremophile MAGs (n = 3,085) containing extremocins. Branch colors denote phyla, while the outer rings indicate habitat type

(Tibet aquifer: Tibet Plateau aquifer; Tibet sediment: Tibet Plateau sediment; T.C. hotspring: Tengchong hotspring; Global deep-sea hydrothermal vent: Deep-sea vent; Other habitats: Antarctica soils, Arctic Sea, Glacier-fed stream, Cold seep and Mariana Trench sediments) and cAMP count for each species. **b.** The main prokaryote phylum proportion in nine habitats. **c.** The phylum-specific density (the count per Giga base pair), diversity and specificity of extremocins. **d.** (i) UMAP reduction of extremocins from bacteria and archaea, and AMPs from DBAASP using sequence similarity matric. (ii) UMAP reduction colored by its category, including EP, REP and DBAASP AMP. **e.** Normalized amino acid frequency of EPs, REPs and AMPs from DBAASP. **f.** Distributions of four peptide physicochemical properties: hydrophobicity, net charge, aliphatic index and normalized nutrient cost. Hydrophobicity influences the peptide interaction with membrane lipids and net charge influences peptides' electrostatic interactions with negatively charged bacterial membranes. The aliphatic index reflects the proportion of aliphatic side chain, with higher indicating the increased thermostability [43]. The nutrient costs normalized by sequence length reflects the energy cost of peptide biosynthesis in nutrient-limited environments.

## In vitro antimicrobial activities of extremocins

We used the following criteria to filter out compounds: (1) sequences longer than 35 amino acid residues, (2) sequences with high sequence similarity to known AMPs from DBAASP and (3) REPs that are present in the reference microbial genomes. As a pre-experiments, we selected and synthesized 14 candidate extremocins based on the predicted score gradient. Two of them showed stronger antimicrobial activity then commercial AMP, LL-37, in vitro against three clinically relevant pathogens, including Escherichia coli, Pseudomonas aeruginosa and Staphylococcus aureus.

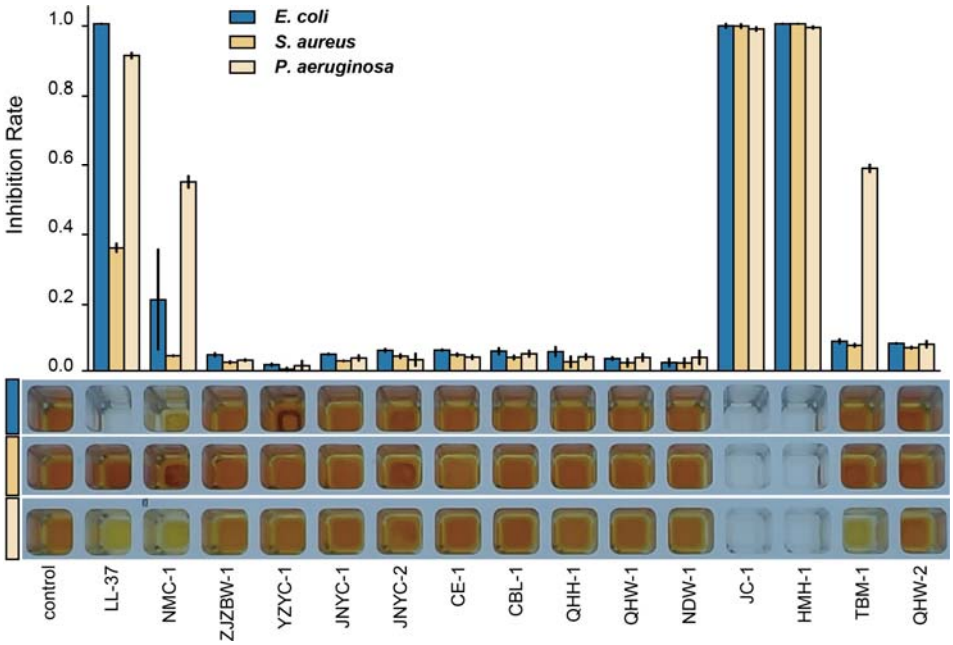


**Figure 3. Inhabitation rate of the candidate AMPs (64 μmol $l^{-1}$) against clinically relevant pathogens.** $10^6$ bacterial cells were incubated at 37 □, Bacterial cultures were incubated at 37 □, and growth was assessed by optical density (OD) at 450 nm after 5 hours of treatment.

## Discussion

In this study, we developed the SEGMA framework to systematically explore antimicrobial peptides (AMPs) from extremophile metagenomes, highlighting the potential of extremophiles, particularly archaeal lineages, as an untapped reservoir of antibiotics. We constructed a global smORF catalog from 60,461 metagenome-assembled genomes (MAGs) spanning 1,406 metagenome samples worldwide, which contains ~4 million smORFs sourced from bacteria and archaea, representing a vast and novel sequence space.

Using SEGMA modules, we identified a total of 3,298 extremocins, assigning to 1,390 AMP families, from 3,084 MAGs. Taxonomically, archaeal phyla exhibited higher extremocin densities, whereas bacterial phyla demonstrated greater diversity and specificity. Focusing on catalog-specific extremocins (EPs) that lack homologs in reference microbial proteins, these sequences represent novel antibiotic classes relative to conventional AMPs from the DBAASP database. Specifically, metagenomes from Tengchong hot spring and the Tibet Plateau contributed the most and also harbored a higher proportion of EPs, underscoring their potential for antibiotic discovery. From the standpoint of physicochemical properties, extremocins were significantly enriched in cationic amino acids, indicating enhanced electrostatic interactions with negatively charged bacterial membranes.

In summary, we demonstrate that the SEGMA framework effectively mines novel antimicrobial peptides from extremophile genomes, revealing a diverse array of extremocins with unique sequence features that position them as viable candidates for addressing global antimicrobial resistance challenges.

## Methods

## Data collection

In this study, we collected two datasets: non-antimicrobial peptides (non-AMPs), antimicrobial peptides (AMPs) for model training.

**Non-AMPs.** The dataset was downloaded from Uniprot (https://www.uniprot.org/ access time: September 2025) to serve as the dataset for supervised fine-tuning of SaProt-650M and SEGMA-Prospect's training. Entries matching following keywords:

‘antimicrobial’, ‘antibiotic’, ‘antiviral’, ‘antifungal’, ‘antibacterial’, ‘archaeasin’, ‘bacteriocin’, ‘effector’, or ‘excreted’ were removed. The peptides longer than 50 amino acids (AAs) or shorter than 10 AAs were also removed. Following redundancy reduction using CD-HIT at a 50% identity threshold, 14,540 peptides were retained. Of which 10,000 were randomly sampled for supervised fine-tuning, with the remainder allocated for GBM training.

**AMPs.** AMP data were collected from four public database: DRAMP, dbAMP, DRAMP and APD as SEGMA-Prospect’s positive dataset (access time: April 2025), and merged into a comprehensive dataset encompassing AMPs from diverse source. The same length filtering and redundancy reduction procedures as for non-AMPs were applied, resulting in 4,394 unique AMPs for GBM training.

## Construction of extremophile smORF catalog

**Large-scale metagenomes.** 60,461 extremophile metagenome-assembled genomes (MAGs) were collected and downloaded from the NCBI Bioproject Browser (https://www.ncbi.nlm.nih.gov/bioproject), European Bioinformatics Institute BioStudy Browser (https://www.ebi.ac.uk/biostudies/), and National Genomics Data Center (https://ngdc.cncb.ac.cn/).

**Taxonomic annotation.** The taxonomic annotation of MAGs was conducted using the Genome Taxonomy Database Toolkit (GTDB-Tk v2 [44]) against the GTDB R207 [45] via the ‘classify_wf’ workflow with default parameters. Then, the phylogenetic trees were inferred with the FastTree [46] based on the multiple sequence alignments.

**smORF prediction and clustering.** Modified Prodigal was applied to predict smORFs from the metagenomes, yielding 3,784,363 sequences with lengths between 11-51 AAs. They were further clustered into protein families using CD-HiT with the following parameters: -c 0.5 -n 2 -T 128 -g 1 -s 0.95 -aL 0.95 -l 5. This approach grouped sequences with ≥50% sequence identity, requiring a minimum coverage of 95% for the shorter sequence, resulting in 2,140,838 protein families.

**Homolog search.** We downloaded all small proteins from SmProt [28] ($n = 449{,}335$), Uniref90 [27] ($n = 2{,}340{,}182$) and GMSC [9] ($n = 108{,}844{,}508$) with length control (aa length < 51). We compared the representative sequences from our extremophile catalog ($n = 2{,}140{,}838$) with these public datasets ($n = 111{,}634{,}025$) using DIAMOND with the ‘-more-sensitive’ mode, keeping significant hits ($E\ value < 10^{-5}$). A total of 396,856 smORFs (18.54%) have homologs in these datasets (SmProt: 3,755; Uniref90: 258,441; GMSC: 309,425).

## Representation of peptides

**Physicochemical property calculation.** To analyze the physicochemical properties of cAMPs from different extreme environments and reference AMPs from the DBAASP database, we used the Python package *Peptides* [47] to calculate the following properties which are commonly associated with peptide antimicrobial activity: hydrophobicity, net charge, isoelectric point, aliphatic index, instability index, Boman index, normalized nutrient costs and energy costs. The net charge was calculated based on *Henderson–Hasselbalch* equation. The Boman index quantifies a peptide's protein-binding affinity by calculating the mean solubility free energy of its constituent amino acids [48].

**Structural information incorporation.** To incorporate structural information into peptide seuqnece representations, ESMFold was used to predict structures from the amino acid sequences. Then, Foldseek was used to tokenize structures by converting atomic coordinates into 3Di tokens, which captrues local geometric relationships and backbone arrangements. Finally, the 3Di tokens were combined with the original amino acid sequences, yielding a hybrid representation that integrates both sequence-level and structure-level information.

**Peptide sequence similarity.** Let $SW(i,j)$ denote Smith - Waterman alignment score between peptide sequences $i$ and $j$, computed via Biopython's pairwise localxx with match score 1, mismatch 0, and no gap penalties. The similarity function is defined as

$$\frac{SW(i,j)}{\sqrt{SW(i,i)\cdot SW(j,j)}},$$

where $SW(i,i) = len(i)$ and $SW(j,j) = len(j)$, or 0.0 if no alignment exists.

## SEGMA architecture

SEGMA comprised of two steps: SEGMA-Prospect and SEGMA-Search, and three models: a Gradient Boosting model (GBM), SaProt-LoRA-1 and SaProt-LoRA-2. The GBM was trained to predict peptide antimicrobial potency based on embeddings generated by SaProt-LoRA-1. Two SaProt-LoRA models were fine-tuned on extremophile smORFs to capture specific evolutionary information.

**SaProt-LoRA.** To enable efficient fine-tuning of the language model with less trainable parameters, we adopted the Low-Rank Adaptation **(**LoRA) strategy. For a given pretrained weight matrix $W_0 \in \mathbb{R}^{d\times k}$, LoRA introduces two low-rank decomposition matrices $A \in \mathbb{R}^{r\times k}$ and $B \in \mathbb{R}^{d\times r}$ to constrain weght updation, where rank $r \ll \min(d,k)$. The forward pass of the adapted layer, given an input $x \in \mathbb{R}^k$, becomes: $h = W_0 x + BAx$, where only weights of $A$ and $B$ are updated via backpropagation and scaled by the quantity $r$ and $\alpha$. During fine-tuning, 25% of residues were randomly masked, the model was trained to reconstruct the masked

residues from the remaining sequence and complete structural context. Based on the strategy, we trained two models: SaProt-LoRA-1 and SaProt-LoRA-2. SaProt-LoRA-1 was trained on extremophile smORFs and Uniprot small proteins to enocde diverse peptides to a unfied embedding space, which enables GBM to predict antimicrobial potency robustly in SEGMA-Prospect module. Furthermore, SaProt-LoRA-2 was trained on extremophile smORFs and AMPs from the DBAASP database to capture antimicrobial motifs and guide peptide mutation in SEGMA-Search module.

**GBM.** The GBM was trained on 4,394 AMPs and 4,540 non-AMPs. Each peptide was encoded by SaProt-LoRA-1 as a 1280-dimensional embedding.

**Training implementation.** The SaProt-LoRA models were implemented in PyTorch, and trained on four NVIDIA A100 GPUs with 80GB memory. The batch size on each GPU was set to 50. The Adam optimizer was applied with an initial learning rate of $5 \times 10^{-5}$. Training was set to run for a maximum of 10 epochs, employing an early stopping strategy to retain the best-performing model. The configuration featured a rank $r = 8$ and alpha $\alpha = 8$ resulting in 0.212% trainable parameters. Besides, the GBM was implemented in Scikit-learn, and trained on Intel(R) Xeon(R) Platinum 8358P CPU. The max iteration was set to 100.

## Quantification of phylum-specific antimicrobial potential

To access phylum-specific antimicrobial potential, we defined and calculated three metrics, including density, diversity and specificity of extremocins across ten main phyla using the following equations:

$$cAMP\ density = \frac{cAMP\ count}{Genome\ size}$$

$$cAMP\ diversity = \frac{count(cAMP\ family)}{count(cAMP)} \cdot \frac{count(Species)}{count(Genome)}$$

$$cAMP\ specificity = \frac{count(unique\ cAMP\ family)}{count(\ cAMP\ family)}$$

## Code availability

The code for SEGMA is available at https://github.com/HUST-NingKang-Lab/SEGMA

## Acknowledge

This work was partially supported by the National Key R&D Program of China (Grant No. 2023YFA1800900, 2021YFA0910500 and 2018YFC0910502) and the National Natural Science Foundation of China (22574058, 32571640).

## Reference

1. Ventola, C.L., *The antibiotic resistance crisis: part 1: causes and threats.* Pharmacy and therapeutics, 2015. **40**(4): p. 277.
2. Oliveira Júnior, N.G., et al., *Antimicrobial peptides: structure, functions and translational applications.* Nature Reviews Microbiology, 2025: p. 1-14.
3. Shu, W.-S. and L.-N. Huang, *Microbial diversity in extreme environments.* Nature Reviews Microbiology, 2022. **20**(4): p. 219-235.
4. Szabó, G., et al., *A genomic view of Earth's biomes.* Nature Reviews Genetics, 2025.
5. Makhalanyane, T.P., *Illuminating microbial defence systems in extreme environments.* Nature Reviews Microbiology, 2023. **21**(10): p. 636-636.
6. Rothschild, L.J. and R.L. Mancinelli, *Life in extreme environments.* Nature, 2001. **409**(6823): p. 1092-1101.
7. Merino, N., et al., *Living at the extremes: extremophiles and the limits of life in a planetary context.* Frontiers in microbiology, 2019. **10**: p. 780.
8. Torres, M.D.T., F. Wan, and C. de la Fuente-Nunez, *Deep learning reveals antibiotics in the archaeal proteome.* Nature Microbiology, 2025. **10**(9): p. 2153-2167.
9. Duan, Y., et al., *A catalog of small proteins from the global microbiome.* Nat Commun, 2024. **15**(1): p. 7563.
10. Wan, F., et al., *Deep-learning-enabled antibiotic discovery through molecular*

*de-extinction.* Nature Biomedical Engineering, 2024. **8**(7): p. 854-871.

11. Huang, J., et al., *Identification of potent antimicrobial peptides via a machine-learning pipeline that mines the entire space of peptide sequences.* Nature Biomedical Engineering, 2023. **7**(6): p. 797-810.

12. Chen, N., et al., *TP-LMMSG: a peptide prediction graph neural network incorporating flexible amino acid property representation.* Briefings in Bioinformatics, 2024. **25**(4): p. bbae308.

13. Bhadra, P., et al., *AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest.* Scientific reports, 2018. **8**(1): p. 1697.

14. Li, W., et al., *Unveiling the evolution of antimicrobial peptides in gut microbes via foundation-model-powered framework.* Cell Reports, 2025. **44**(6).

15. Wang, X.F., et al., *ProT-Diff: A Modularized and Efficient Strategy for De Novo Generation of Antimicrobial Peptide Sequences by Integrating Protein Language and Diffusion Models.* Advanced Science, 2024. **11**(43): p. 2406305.

16. Luo, J., et al., *iMFP-LG: identify novel multi-functional peptides using protein language models and graph-based deep learning.* Genomics, Proteomics and Bioinformatics, 2024. **22**(6): p. qzae084.

17. Ma, Y., et al., *Identification of antimicrobial peptides from the human gut microbiome using deep learning.* Nature Biotechnology, 2022. **40**(6): p. 921-931.

18. Wang, Y., et al., *A generative artificial intelligence approach for the discovery of antimicrobial peptides against multidrug-resistant bacteria.* Nature Microbiology, 2025:

p. 1-16.

19. Dong, R., Q. Cao, and C. Song, *Painting Peptides With Antimicrobial Potency Through Deep Reinforcement Learning.* Advanced Science, 2025: p. e06332.

20. Gao, H., et al., *DLFea4AMPGen de novo design of antimicrobial peptides by integrating features learned from deep learning models.* Nature Communications, 2025. **16**(1): p. 9134.

21. Wang, Y., et al., *Controllable Generation of Pathogen-Specific Antimicrobial Peptides Through Knowledge-Aware Prompt Diffusion Model.* Advanced Science, 2025: p. e07457.

22. Wang, J., et al., *Discovery of antimicrobial peptides with notable antibacterial potency by an LLM-based foundation model.* Science advances, 2025. **11**(10): p. eads8932.

23. Dong, R., et al., *Exploring the repository of de novo-designed bifunctional antimicrobial peptides through deep learning.* eLife, 2025. **13**: p. RP97330.

24. Zhang, H., et al., *Machine learning and genetic algorithm-guided directed evolution for the development of antimicrobial peptides.* Journal of Advanced Research, 2025. **68**: p. 415-428.

25. Cheng, M., et al., *A genome and gene catalog of the aquatic microbiomes of the Tibetan Plateau.* Nat Commun, 2024. **15**(1): p. 1438.

26. Ding, L., et al., *Timing and mechanisms of Tibetan Plateau uplift.* Nature Reviews Earth & Environment, 2022. **3**(10): p. 652-667.

27. Suzek, B.E., et al., *UniRef: comprehensive and non-redundant UniProt reference clusters.* Bioinformatics, 2007. **23**(10): p. 1282-1288.

28. Hao, Y., et al., *SmProt: a database of small proteins encoded by annotated coding and non-coding RNA loci.* Briefings in Bioinformatics, 2017. **19**(4): p. 636-643.

29. Lin, Z., et al., *Evolutionary-scale prediction of atomic-level protein structure with a language model.* Science, 2023. **379**(6637): p. 1123-1130.

30. van Kempen, M., et al., *Fast and accurate protein structure search with Foldseek.* Nat Biotechnol, 2024. **42**(2): p. 243-246.

31. Hu, E.J., et al., *Lora: Low-rank adaptation of large language models.* ICLR, 2022. **1**(2): p. 3.

32. Schmirler, R., M. Heinzinger, and B. Rost, *Fine-tuning protein language models boosts predictions across diverse tasks.* Nature Communications, 2024. **15**(1): p. 7407.

33. Sledzieski, S., et al., *Democratizing protein language models with parameter-efficient fine-tuning.* Proceedings of the National Academy of Sciences, 2024. **121**(26): p. e2405840121.

34. Parks, D.H., et al., *A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life.* Nature biotechnology, 2018. **36**(10): p. 996-1004.

35. Hug, L.A., et al., *A new view of the tree of life.* Nature microbiology, 2016. **1**(5): p. 1-6.

36. Castelle, C.J. and J.F. Banfield, *Major new microbial groups expand diversity and alter our understanding of the tree of life.* Cell, 2018. **172**(6): p. 1181-1197.

37. Cunin, R., et al., *Biosynthesis and metabolism of arginine in bacteria.* Microbiological reviews, 1986. **50**(3): p. 314-352.

38. Xu, Y., et al., *Evolution of arginine biosynthesis in the bacterial domain: novel*

*gene-enzyme relationships from psychrophilic Moritella strains (Vibrionaceae) and evolutionary significance of N-α-acetyl ornithinase.* Journal of bacteriology, 2000. **182**(6): p. 1609-1615.

39. Hancock, R.E. and R. Lehrer, *Cationic peptides: a new source of antibiotics.* Trends in biotechnology, 1998. **16**(2): p. 82-88.

40. Yeung, A.T.Y., S.L. Gellatly, and R.E.W. Hancock, *Multifunctional cationic host defence peptides and their clinical applications.* Cellular and Molecular Life Sciences, 2011. **68**(13): p. 2161-2176.

41. Cantalapiedra, C.P., et al., *eggNOG-mapper v2: functional annotation, orthology assignments, and domain prediction at the metagenomic scale.* Molecular biology and evolution, 2021. **38**(12): p. 5825-5829.

42. Hurtado-Rios, J.J., et al., *Ribosomes: the new role of ribosomal proteins as natural antimicrobials.* International journal of molecular sciences, 2022. **23**(16): p. 9123.

43. IKAI, A., *Thermostability and Aliphatic Index of Globular Proteins.* The Journal of Biochemistry, 1980. **88**(6): p. 1895-1898.

44. Chaumeil, P.-A., et al., *GTDB-Tk v2: memory friendly classification with the genome taxonomy database.* Bioinformatics, 2022. **38**(23): p. 5315-5316.

45. Parks, D.H., et al., *A complete domain-to-species taxonomy for Bacteria and Archaea.* Nature biotechnology, 2020. **38**(9): p. 1079-1086.

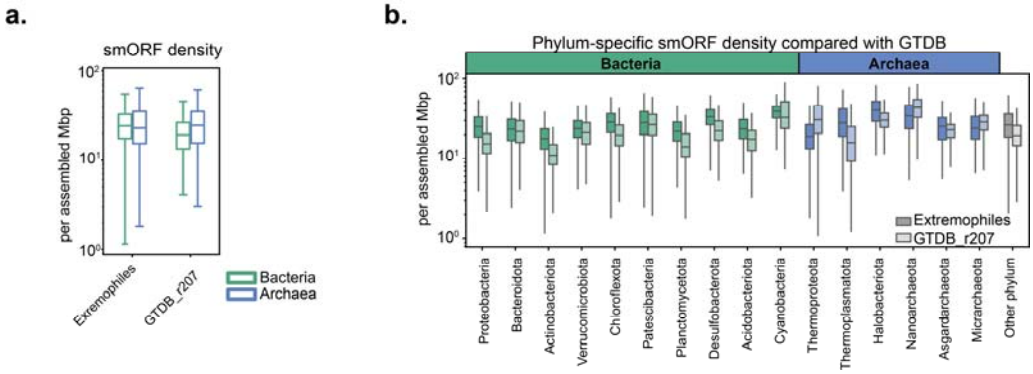
46. Price, M.N., P.S. Dehal, and A.P. Arkin, *FastTree 2–approximately maximum-likelihood trees for large alignments.* PloS one, 2010. **5**(3): p. e9490.

47. Osorio, D., P. Rondón-Villarreal, and R. Torres, *Peptides: a package for data mining*

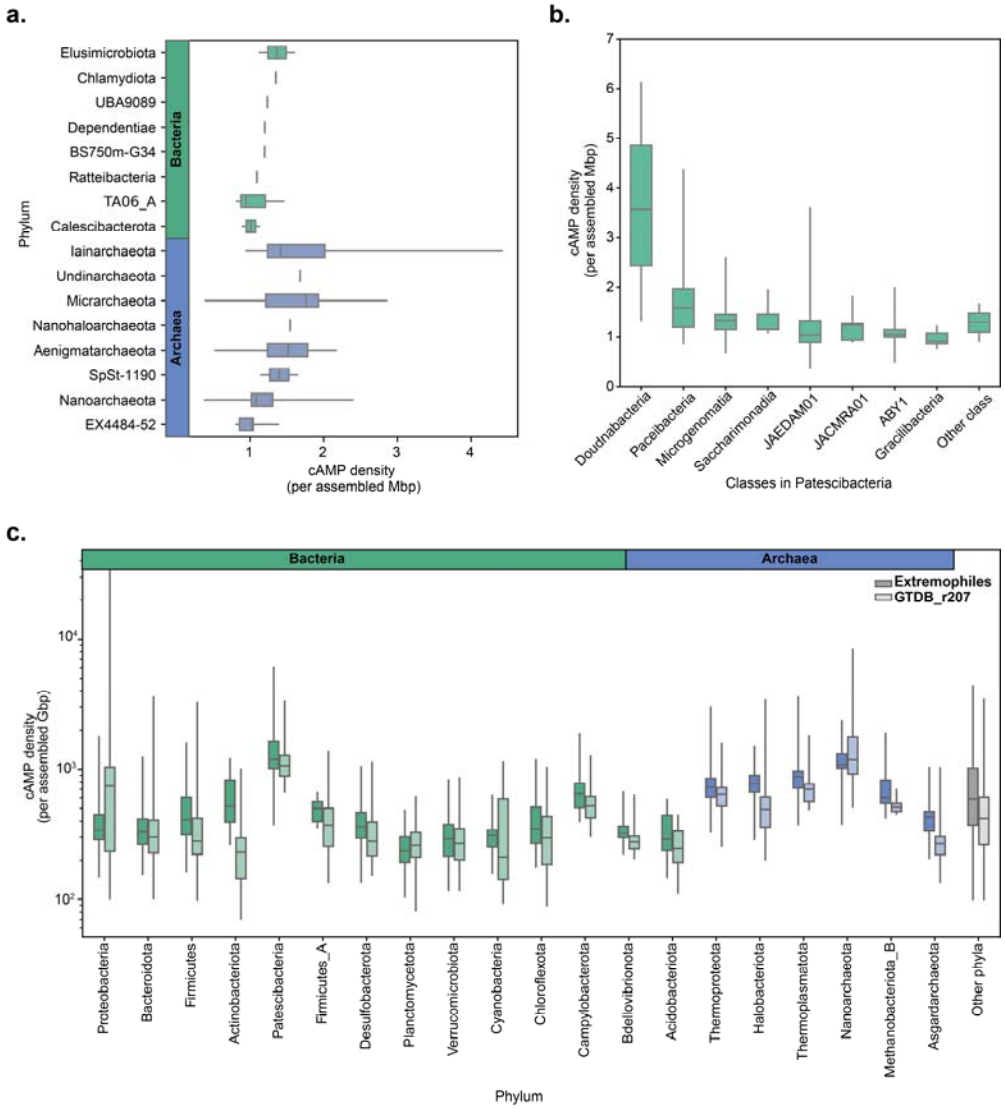
*of antimicrobial peptides.* 2015.

48. Radzicka, A. and R. Wolfenden, *Comparing the polarities of the amino acids: side-chain distribution coefficients between the vapor phase, cyclohexane, 1-octanol, and neutral aqueous solution.* Biochemistry, 1988. **27**(5): p. 1664-1670.

# Supplementary information

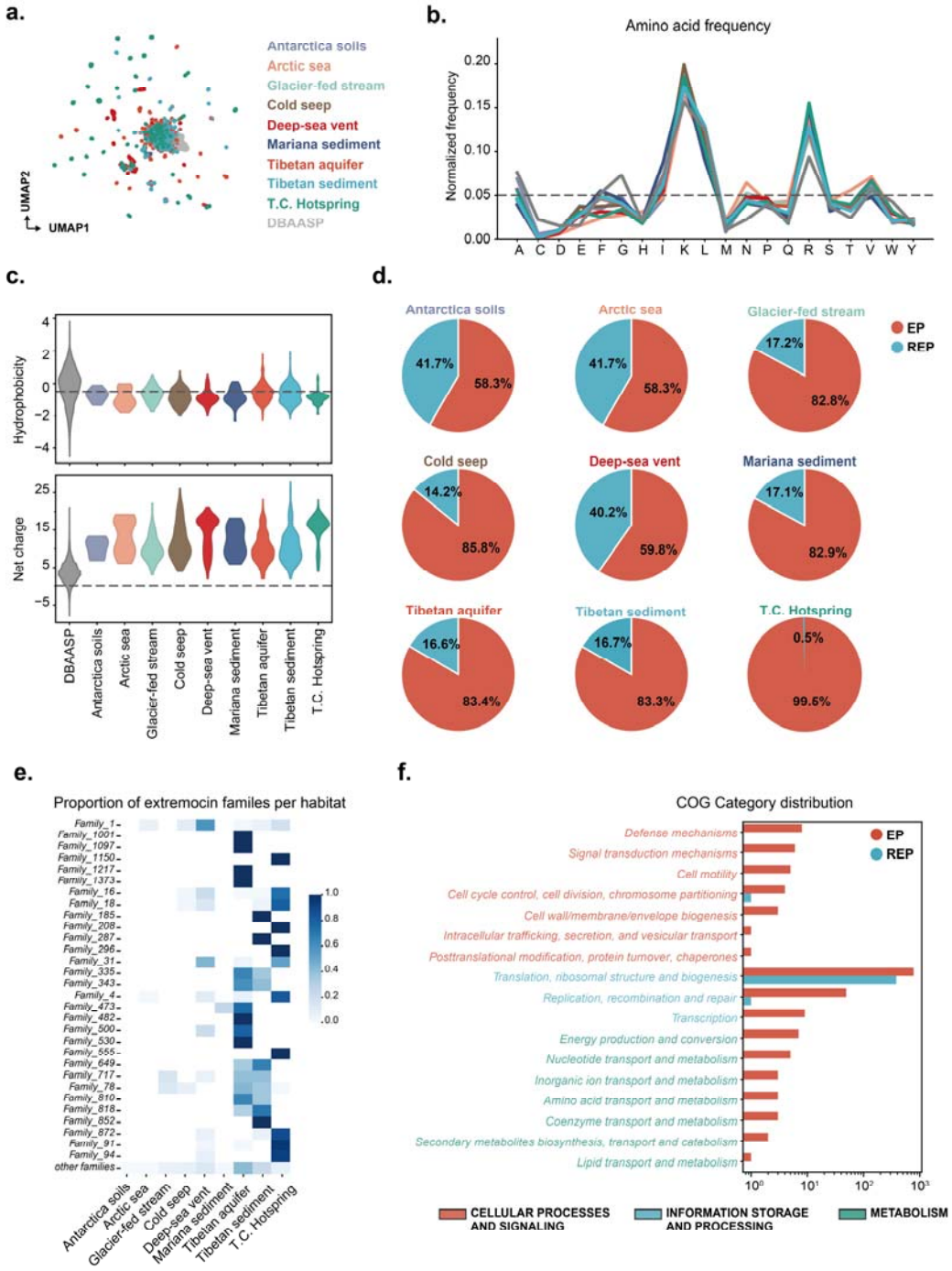


**Supplementary Figure 1. Extremophile smORF density compared with reference microbial genomes. (a)** Bacteria from extreme environments harbor higher smORF density (the number of smORFs per megabase pairs, Mbp) than genomes from GTDB ($P_{Mann} < 0.0001$). The density of archaeal phyla is higher than bacterial ones in GTDB ($P_{Mann} < 0.0001$). But in extreme environments, bacteria and archaea exhibit similar density ($P_{Mann} = 0.0026$). **(b)** In general, main bacteria phyla ($n > 1000$) and archaea phyla ($n > 200$) in the extremophile catalog showed higher smORF density than those in GTDB.



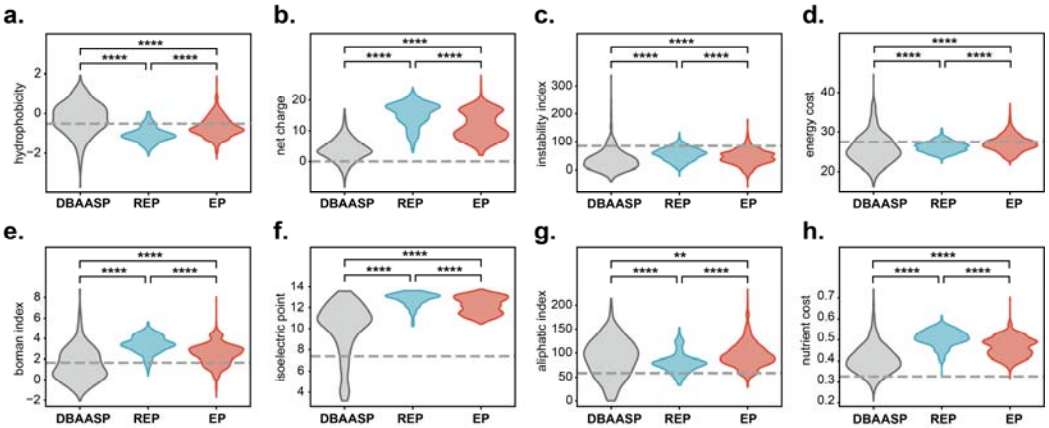

**Supplementary Figure 2. (a)** Extremocin density of non-dominant ($n < 64$) phyla with average density exceeding 1,000 per assembled Mbp. Within these non-dominant phyla, archaea exhibit higher density than bacteria. Elusimicrobiota displays the highest density due to its highly reduced genomes (average assembled genome length: 1,066,299 bps). **(b)** Extremocin density across different classes within Patescibacteria. Doudnabacteria exhibits the highest density. **(c)** Extremocin density compared with

AMP density of reference microbial genomes. In the most phyla, extremophiles harbor higher density.

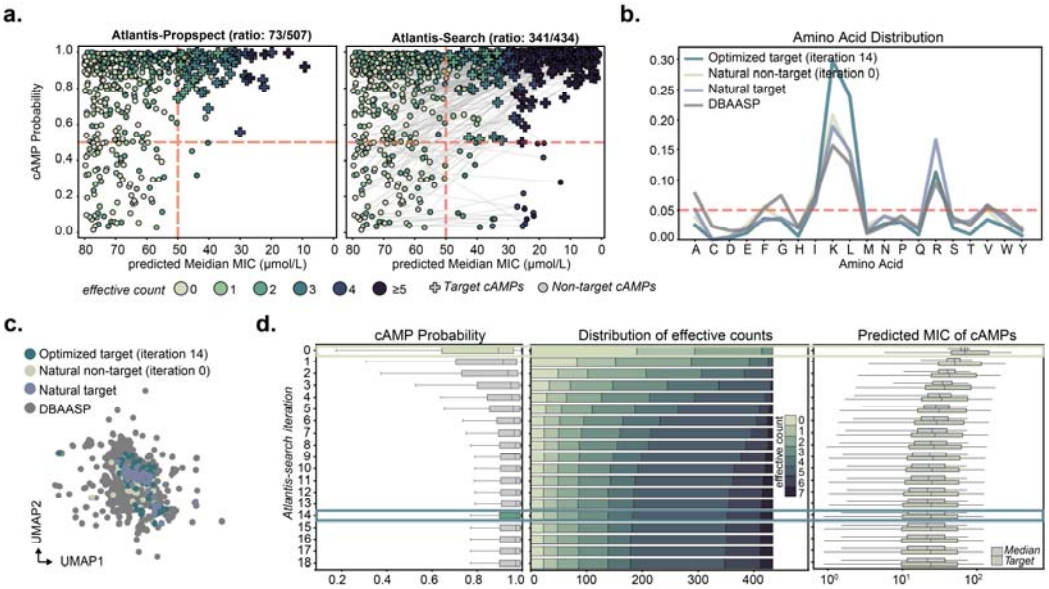



**Supplementary Figure 3. Habitat-specific properties, proportions and annotations of extremocins. (a)** UMAP reduction of habitat-specific extremocins and AMPs from DBAASP using sequence similarity matric. **(b)** Normalized amino acid frequencies. **(c)** Physiochemical features, including hydrophobicity and net charge. **(d)** EP and REP proportion of extremocins per habitats. **(e)** Proportion of extremocin

families per habitat. **(f)** The COG distribution of EPs and REPs.




**Supplementary Figure 4. Physiochemical features of EPs and REPs detected from extremophile metagenomes compared to AMPs from DBAASP. (a)** Hydrophobicity, **(b)** net charge, **(e)** Boman index and **(f)** isoelectric point directly reflect peptide propensity to bind with bacterial membrane lipids. **(c)** Instability and **(g)** aliphatic index reflect peptide stability. **(d, h)** Energy and nutrient costs normalized by sequence length are both used to estimate the energy cost required to biosynthesize a peptide based on genome-scale metabolic modeling in *Escherichia coli*. Statistical significance was calculated using two-tailed Mann-Whitney test.




**Supplementary Figure 5. Refinement and optimization of candidate AMPs.**