A Global Discovery of Antimicrobial Peptides in Deep-Sea Microbiomes Driven by an ESM-2 and Transformer-based Dual-Engine Framework

Bairun Chen[1], Xinyi Mou[1], Zhuoxuan Song[1], Huaying Lin[2], Yu Zhang[2*], Jing Li[1*]

[1] Department of Bioinformatics and Biostatistics, School of Life Sciences and Biotechnology, Shanghai Jiao Tong University, Shanghai 200240, Shanghai, P. R. China.

[2] Shanghai Key Laboratory of Polar Life and Environment Sciences, School of Oceanography, Shanghai Jiao Tong University, Shanghai 200240, Shanghai, P. R. China.

* Corresponding author: Jing Li (jing.li@sjtu.edu.cn), Yu Zhang (zhang.yusjtu@sjtu.edu.cn)

Abstract

The escalating global threat of multidrug-resistant pathogens underscores the urgent demand for innovative strategies in antimicrobial peptide (AMP) discovery. Notably, deep-sea-related data resources remain underexplored, despite their potential as valuable sources of novel AMPs. Current AMP prediction methods, however, are limited by dataset biases such as sequence length imbalance between AMPs and non-AMPs, N-terminal methionine artifacts in non-AMPs, and microbial origin specificity. To overcome these constraints, we developed a dual-engine predictor named XAMP, which integrates two complementary architectures: XAMP-E, built on ESM-2 for high-accuracy feature representation, and XAMP-T, built on one-layer Transformer for accelerating large-scale screening. This dual-engine design ensures both robust feature learning and enhanced generalization capability. By constructing length-balanced datasets, removing N-terminal methionine from non-AMPs, and training microbial-specific variants, XAMP achieved a median area under the receiver-operating characteristic curve (AUC) of 0.972, representing an approximately 10% improvement over state-of-the-art predictors. Using an integrated AMP mining pipeline incorporating these models, we screened large-scale deep-sea metagenomic data and identified 2,355 promising AMP candidates. This study establishes a robust deep-learning framework that facilitates targeted discovery of bioactive peptides from extremophile microbiomes through systematic multi-omics validation.

KEY WORDS: Antimicrobial peptides, Transformer-based model, ESM-2, Metagenomes, Metaproteomics, Deep-sea microbiome, Microbial smORFs

Introduction

Antibiotics have been foundational to modern medicine[1], but their overuse and the horizontal transmission of resistance genes have fueled the global crisis of multi-drug resistant (MDR) bacteria[2]. By 2050, MDR infections are projected to cause over 10 million annual deaths, surpassing cancer as a leading cause of mortality[3]. This crisis demands alternative therapeutics[2], with antimicrobial peptides (AMPs) emerging as promising candidates due to their broad-spectrum activity and unique mechanism of action—targeting bacterial membranes rather than specific genetic pathways, thus minimizing resistance development[4][5].

Recent breakthroughs in metagenomic analysis and artificial intelligence (AI) have accelerated AMP discovery. Marcel, a random forest-based AMP prediction tool, integrates six local and sixteen global features to enable end-to-end prediction directly from (meta)genomic raw data[6]. The deep learning framework c_AMPs-prediction, employs an architecture combining Long Short-Term Memory (LSTM), attention mechanisms, and Bidirectional Encoder Representations from Transformers (BERT) for identifying AMPs within human gut microbiota[7]. iAMPCN, a Convolutional Neural Network (CNN)-based model, achieves accurate identification of 22 functional activities by integrating four distinct types of sequence features[8]. APG leverages the ESM-2 protein language model as a discriminator to differentiate between real and fake peptides generated by its generator[9]. These tools have enabled screening with high efficiency, but there are still many critical biases persist: (i). Length distribution imbalance: Most predictors use UniProt-derived negative samples, which lack short proteins, leading to spurious negative correlations between prediction probability and length. Also, APG's short length focus on 11-30 aa underscore the need for broadly applicable predictors. (ii). N-terminal methionine (Met) artifacts: Mature proteins often undergo N-Met excision, but 92.7% of non-AMPs in c_AMPs's trainset retain N-Met, distorting feature spaces. (iii). Microbial origin specificity: Existing models lack optimization for microbial AMPs, which dominate deep-sea and extremophilic environments but differ in physicochemical properties (e.g., charge and hydrophobicity) from non-microbial AMPs.

While these methodological challenges persist, AI-driven mining of metagenomic data has yielded substantial breakthroughs in discovering AMPs from diverse biological sources. From the human gut microbiome, Ma et al. identified 2,349 candidate AMPs using an integrated NLP pipeline, with experimental validation confirming activity in 181 of 216 synthesized peptides, showing efficacy even against drug-resistant lung infections in mice[7]. Similarly, Xu et al. uncovered a large reservoir of AMPs from engineered environments such as sludge, identifying 27 candidates of which 21 exhibited antibacterial activity[10]. On a global scale, Santos et al. established the AMPSphere

resource by screening over 150,000 metagenomes and genomes, predicting 863,498 non-redundant AMPs and validating 79 out of 100 tested peptides against clinically relevant pathogens[11]. Beyond contemporary ecosystems, pioneering studies have also explored unconventional sources: Guan et al. mined animal venoms, identifying 53 active peptides out of 58 tested[12], while Torres et al. uncovered archaeasins from archaeal proteomes, with 93% of synthesized candidates showing potent antibacterial activity in vitro and in vivo[13].

Despite these advances, one major biome remains largely untapped: the deep-sea microbiome. Characterized by extreme conditions and unique evolutionary pressures, deep-sea microorganisms represent an exceptional resource for discovering AMPs with novel mechanisms and structures. However, deep-sea-derived AMP resources have been significantly under-explored, partly due to technical challenges in sampling, sequencing, and computational prediction. To bridge this gap, we present a dedicated computational and experimental framework for mining deep-sea microbial AMPs, leveraging a dual-engine deep learning strategy to maximize discovery accuracy and biological relevance. This work aims to unlock the therapeutic potential hidden within the microbial dark matter of the ocean's depths, offering new avenues to combat multidrug-resistant pathogens.

Results

AI-Driven Discovery of Deep-Sea Microbiomes AMPs

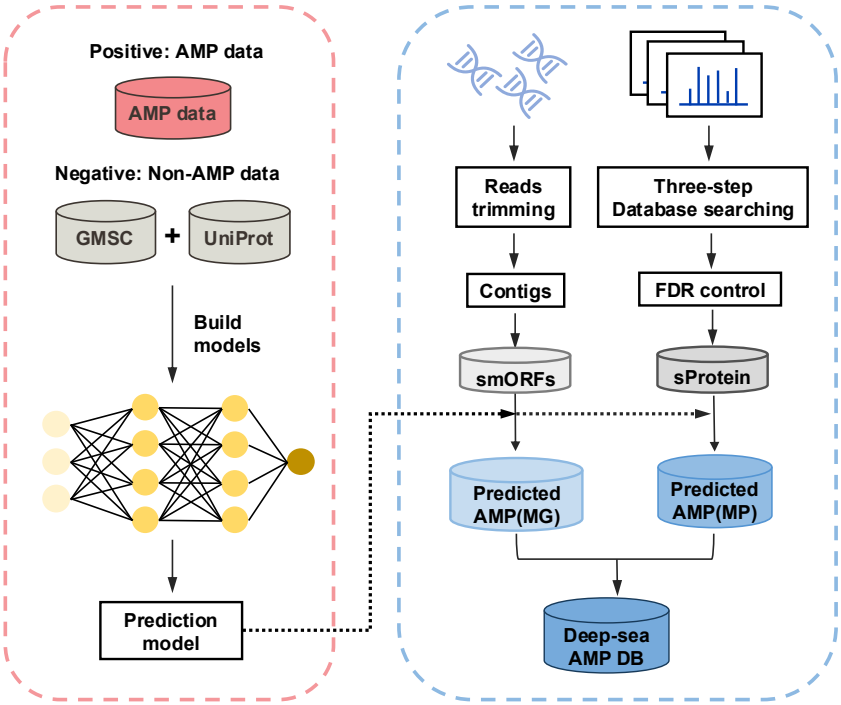


Figure 1. Schematic workflow for mining AMPs from deep-sea microbiomes. *Beginning with curated sequence datasets, we constructed training and test sets. Subsequently, we developed and optimized the antimicrobial peptide (AMP) prediction model, referred to as XAMP, establishing the AMP identification pipeline (left). We then employed this pipeline to screen deep-sea metagenomic and metaproteomic datasets for potential AMPs (right).*

To systematically mining novel AMPs from the deep-sea microbiome, we established a comprehensive research workflow encompassing computational prediction and experimental validation (Figure 1). This process commenced with the construction of a high-quality benchmark dataset, integrating positive AMP sequences from iAMPCN[8] and negative samples (non-AMPs) from the global microbial smORFs catalog (GMSC)[14] and UniProt. Subsequent rigorous data curation addressed prevalent biases, including the balancing of sequence length distribution and correction of N-terminal methionine artifacts. This was followed by the development and optimization of the dual-engine AMP prediction model, XAMP, thereby establishing the core computational identification pipeline. The application of this pipeline to large-scale deep-sea metagenomic and metaproteomic datasets allowed for the screening and identification of high-confidence candidate AMPs.

Quantifying Critical Trainset Artifacts in Existing AMP Prediction Models

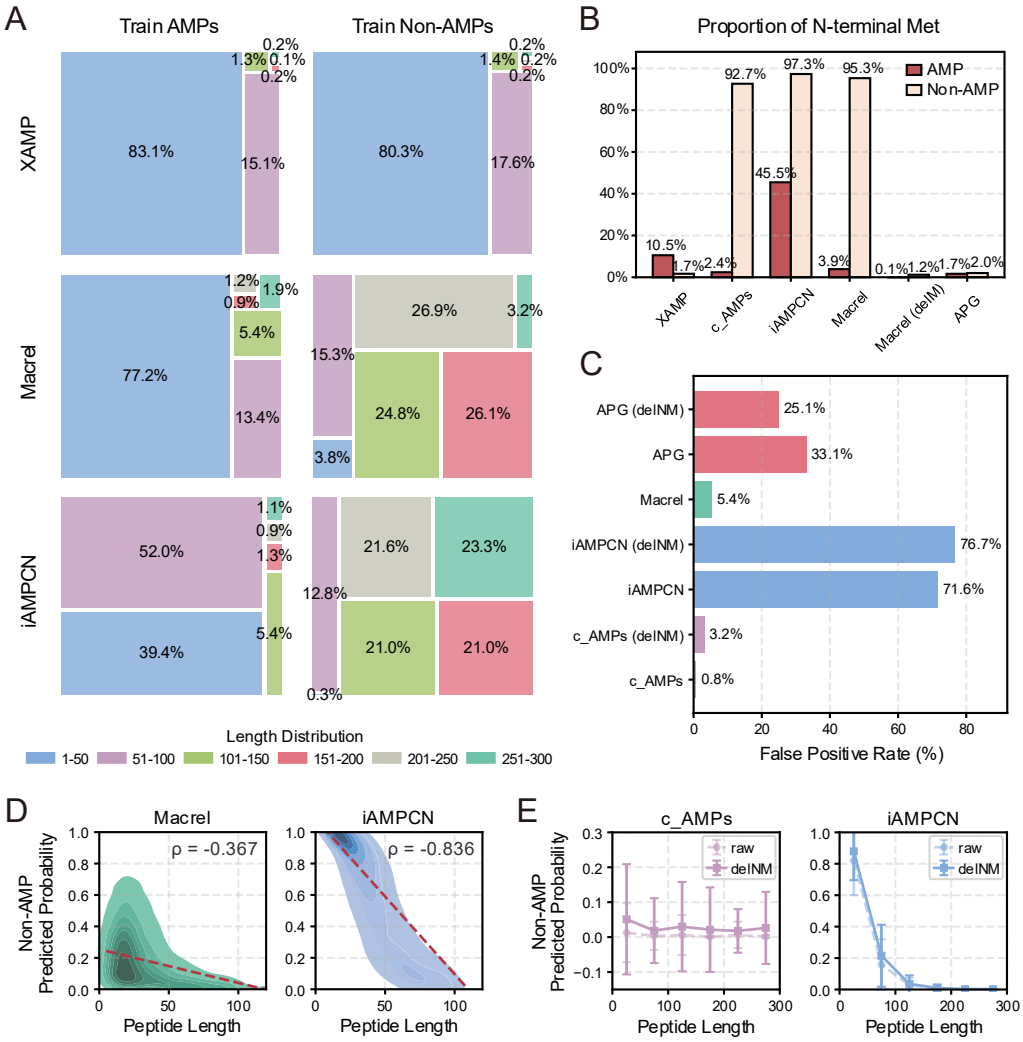


Figure 2. Systematic trainset data biases impair existing AMP prediction reliability.

(A). Length distribution disparities between AMPs and non-AMPs in trainsets of major predictors.

(B). Proportion of N-terminal methionine (N-Met) in AMPs and non-AMPs across trainsets.

(C). False positive rate (FPR) comparison on our debiased testset. "delNM" indicates using N-Met-removed non-AMPs as model inputs.

(D). Spurious negative correlations between prediction probability and peptide length in Macrel (ρ = -0.367) and iAMPCN (ρ = -0.836) on our debiased testset.

(E). Prediction probability of non-AMPs by c_AMPs and iAMPCN before/after N-Met deletion on our debiased testset.

The reliability of existing AMP prediction models is substantially compromised by systematic biases present in their trainsets. First, a pronounced length distribution disparity exists between AMPs and non-AMPs in their trainsets (Figure 2A). For example, less than 4% of non-AMPs are shorter than 50 amino acids in Macrel's dataset, compared to over 77% of AMPs, leading to a spurious negative correlation between prediction score and peptide length (Figure 2D). Second, the prevalence of N-terminal methionine (N-Met) in non-AMP sequences creates an artifact that inflates model performance illusorily. Since most non-AMPs in trainsets retain this translation-initiation residue (92.7% in c_AMPs; 97.3% in iAMPCN; Figure 2B)—unlike mature proteins in vivo[15]—models learn to misuse its presence as a simple marker for negative classification. This results in artificially optimistic performance during validation, which is exposed as a sharp rise in false positive rates when models are tested on debiased data with N-Met removed (Figure 2C, E).

To overcome these limitations, we constructed a rigorously curated benchmark dataset by balancing AMP/non-AMP length distributions and removing N-Met initiation residues from non-AMPs (Figure 2A, B; see *Methods*). The resulting dataset contains more complex sequence characteristics that challenge conventional model architectures, motivating the development of our dual-engine framework, XAMP. By integrating a fine-tuned ESM-2 model for residue-level semantic embedding and a Transformer module for long-range context modeling, XAMP improves the generalization capacity for unbiased AMP recognition.

Development of XAMP for mining AMPs in deep-sea microorganisms

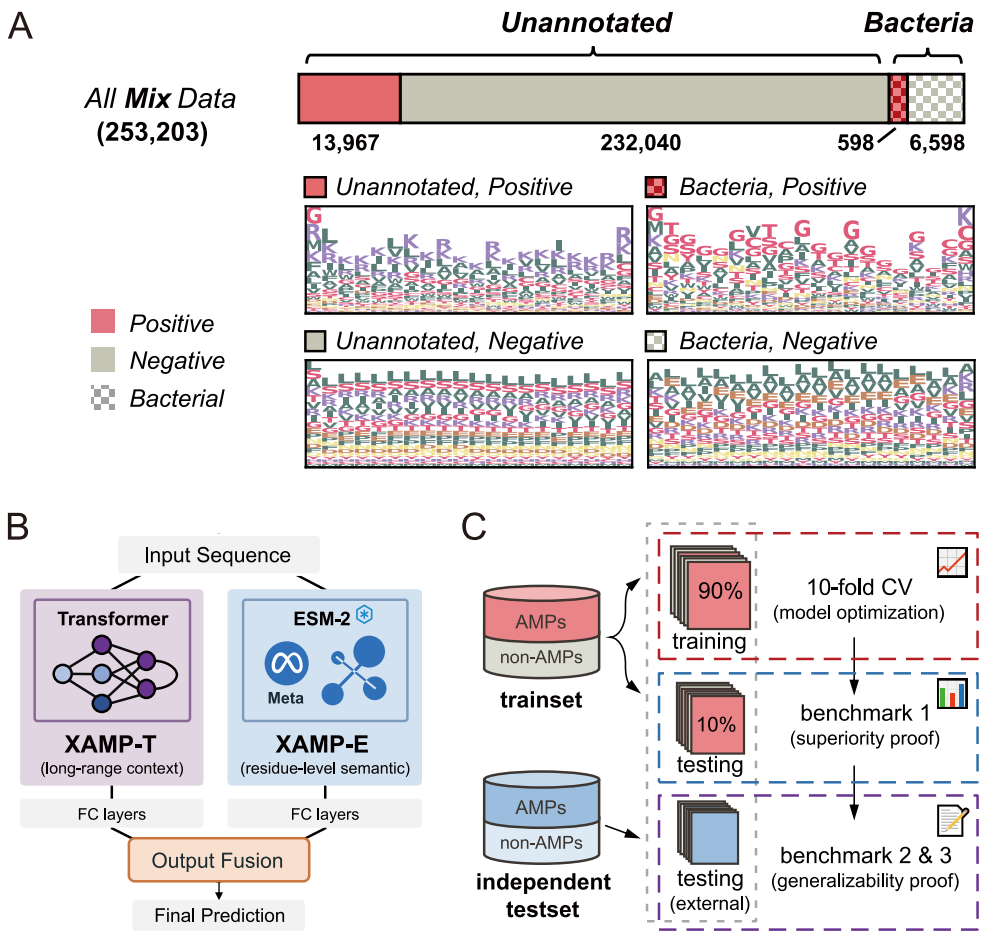


Figure 3. Integrated framework for AMP prediction and analysis.

(A). Data composition and motif characterization. Bar plot displays the distribution of sequences in the 'Mix' dataset. Sequence motifs are shown for N- and C-terminal regions.

(B). Dual-engine architecture of XAMP. The framework integrates two complementary modules: (i) XAMP-T, a Transformer-based encoder for end-to-end sequence classification, and (ii) XAMP-E, which leverages pre-trained ESM-2 embeddings for enhanced feature representation.

(C). Data partitioning strategy. The Trainset was split into training and testing sets, with training employing 10-fold cross-validation. Model performance was evaluated on both internal and external independent testsets.

We constructed a comprehensive benchmark dataset ("Mix") totaling 253,203 sequences, which included 246,007 of unannotated origin (13,967 AMPs; 232,040 non-AMPs) and 7,196 of bacterial origin (598 AMPs; 6,598 non-AMPs) (Figure 3A). Analysis of N- and C-terminal motifs revealed distinct compositional patterns: AMPs of unannotated origin were enriched with positively charged lysine (K) and arginine (R) residues, whereas bacterial AMPs were predominantly composed of

147 glycine (G). Non-AMPs were generally characterized by hydrophobic and polar residues (e.g.,
148 leucine (L), alanine (A), and serine (S)), consistent with background frequencies in source databases.

149 To effectively leverage this data, we developed XAMP, an integrated dual-engine prediction
150 framework (Figure 3B). This architecture combines two complementary approaches: (i) XAMP-T,
151 an end-to-end classifier built on a Transformer encoder followed by fully-connected (FC) layers for
152 deep feature extraction and classification; and (ii) XAMP-E, which augments sequence
153 representation by incorporating embeddings from a pre-trained ESM-2 model, which are then
154 processed through dedicated FC layers. This dual-model strategy provides two application strategies
155 to balance precision and efficiency: using the integrated model that selects the lower prediction
156 probability from both engines to minimize false positives, or employing the faster XAMP-T
157 individually for rapid large-scale screening. This design ensures optimal performance while
158 retaining flexibility for diverse scenarios.

159 Following dataset deduplication, sequences were partitioned into a training (90%) and an internal
160 test set (10%) (Figure 3C). Model hyperparameters were optimized via 10-fold cross-validation,
161 and final performance was rigorously assessed on both the internal test set and an independent
162 external validation set to confirm robustness and generalizability. Detailed dataset statistics and
163 methodological procedures are provided in Supplementary Table 1 and the Methods section,
164 respectively.

The dual-engine XAMP framework enables accurate and efficient AMP identification

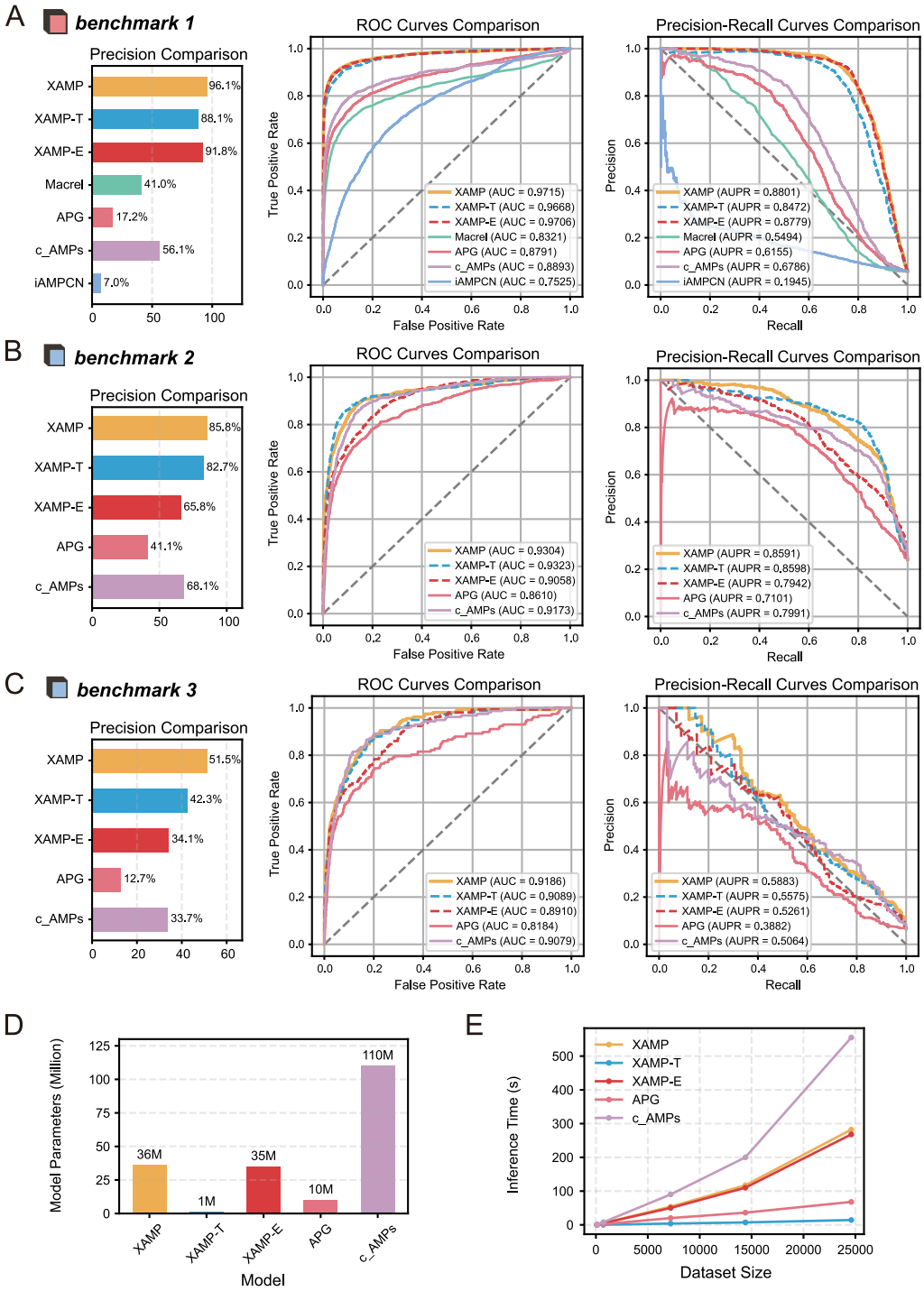


Figure 4. Benchmarking results of XAMP.

(A) Predictive performance on testset (Mix).

(B) Predictive performance on independent testsets (Xiao 1).

(C) Predictive performance on independent testsets (Xiao 2).

(D) Model parameter counts.

(E) Inference speed on datasets of different sizes.

We first assessed the impact of training data through an ablation study. The superior performance of models trained on the "Mix" dataset highlighted the importance of broad data integration for generalization (Supplementary Table 2). Consequently, all subsequent analyses used the ***mix*** models.

In benchmarking, XAMP demonstrated superior predictive accuracy, improving precision by 40.0%-89.1%, AUC by 8.22%-21.9% and AUPRC by 20.2%-68.6% over baseline models on testset (Figure 4A). This advantage held on external, rigorously designed testsets (Figure 4B-C), confirming excellent generalization.

Beyond predictive accuracy, a critical advantage of the XAMP framework lies in its flexible and efficient design. XAMP-T required only 1 million parameters (Figure 4D) and operated 5-40 times faster than other deep learning models across datasets of varying sizes (Figure 4E). This combination of the high-performance, feature-rich XAMP-E engine and the ultra-efficient XAMP-T engine makes the framework uniquely suited for a wide range of scenarios, from detailed analysis to large-scale screening.

Supplementary Table 1. Overview of trainset, testset and independent testset.

Dataset	Positive	Negative	Total	*Remarks
Bacteria (trainset)	526	5,950	6,476	model suffix: "_bt"
Bacteria (testset)	72	648	720	
Unannotated (trainset)	12,585	208,820	221,405	model suffix: "_un"
Unannotated (testset)	1,382	23,220	24,602	
Mix (trainset)	13,111	214,770	227,881	model suffix: "_mix"
Mix (testset)	1,454	23,868	25,322	benchmark 1
Xiao 1 (independent testset)	757	2,386	3,143	benchmark 2: removed the Xiao et al.'s benchmark dataset's AMPs existed in Unannotated (trainset)
Xiao 2 (independent testset)	157	2,206	2,363	benchmark 3: removed the Xiao et al.'s benchmark dataset's AMPs existed in Mix (trainset)

Supplementary Table 2. Model performance ablation study on different training datasets based on the Mix dataset (testset).

Model	Accuracy	Precision	AUC	AUPRC	FPR
XAMP_mix (Ours)	0.9795	**0.9606**	**0.9715**	**0.8801**	**0.0017**
XAMP-T_mix (Ours)	0.9784	0.8807	0.9668	0.8472	0.0059
XAMP-E_mix (Ours)	**0.9822**	0.9176	**0.9706**	**0.8779**	0.0041
XAMP_un (Ours)	0.9798	**0.9600**	0.9616	0.8755	**0.0017**
XAMP-T_un (Ours)	0.9794	0.8802	0.9584	0.8588	0.0062
XAMP-E_un (Ours)	**0.9808**	0.8936	0.9675	0.8683	0.0055
XAMP_bt (Ours)	0.9226	0.2900	0.7770	0.2272	0.0359
XAMP-T_bt (Ours)	0.9052	0.2306	0.7420	0.1900	0.0566
XAMP-E_bt (Ours)	0.9154	0.3341	0.8025	0.3745	0.0578
iAMPCN	0.2738	0.0701	0.7525	0.1945	0.7673
c_AMPs	0.9510	0.5610	0.8893	0.6786	0.0321
Macrel	0.9271	0.4097	0.8321	0.5494	0.0537
APG	0.7551	0.1722	0.8946	0.6248	0.2511

Model Reveals Physicochemical Basis of AMP Antibacterial Activity

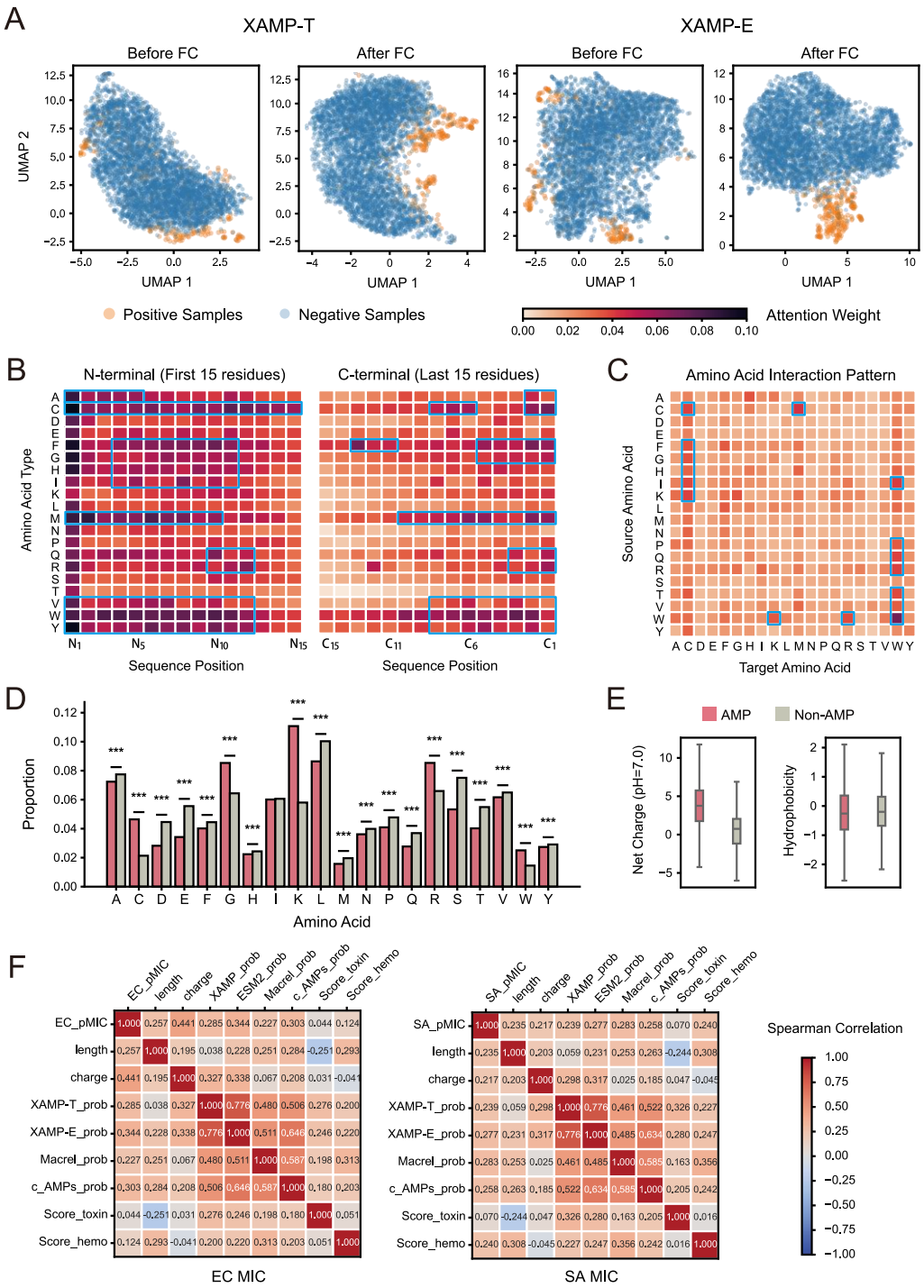


Figure 5. Interpretable features underlying AMP recognition by XAMP.

(A) UMAP visualization shows effective separation of AMPs (orange) from non-AMPs (blue) in the latent space of XAMP-T and XAMP-E, which is refined through the network.

(B) Sequence-level attention in XAMP-T highlights critical residues at N- and C-termini.

(C) Residue interaction matrix reveals cooperative patterns (e.g., W-K/R).

(D, E) Comparative analysis of amino acid composition (D) and key physicochemical properties (E) between AMPs and non-AMPs.

(F) Correlation analysis of model scores with antimicrobial potency (pMIC), peptide properties, and safety metrics.

Interpretability analyses show that XAMP's predictions are grounded in sound physicochemical principles. UMAP visualization confirms that both XAMP-T and XAMP-E effectively separate AMPs from non-AMPs, with separation sharpening after the final layers, indicating refined feature learning (Figure 5A).

Attention analysis in XAMP-T reveals a focus on N/C-terminal residues and preferences for cationic (K/R) and hydrophobic (A/I/V/F/W/Y) amino acids, alongside residue-residue cooperativity (e.g., W-K/R) (Figure 5B, C). These learned patterns align with the significant physicochemical differences between AMPs and non-AMPs[16], notably a much higher net charge (+3.9 vs. +0.5) and distinct amino acid enrichment, despite similar overall hydrophobicity (Figure 5D, E).

While XAMP is a classifier, its prediction score shows a moderate correlation with E. coli pMIC (Spearman's r=0.344) from BERT-AmPEP60[17], exceeding other models (Figure 5F). More critically, the moderate correlations between model scores and toxicity/hemolysis metrics (r=0.20–0.33) highlight that potent AMP candidates may carry safety risks, underscoring the necessity of integrating safety screening after activity prediction.

Mining and characterization of deep-sea antimicrobial peptides

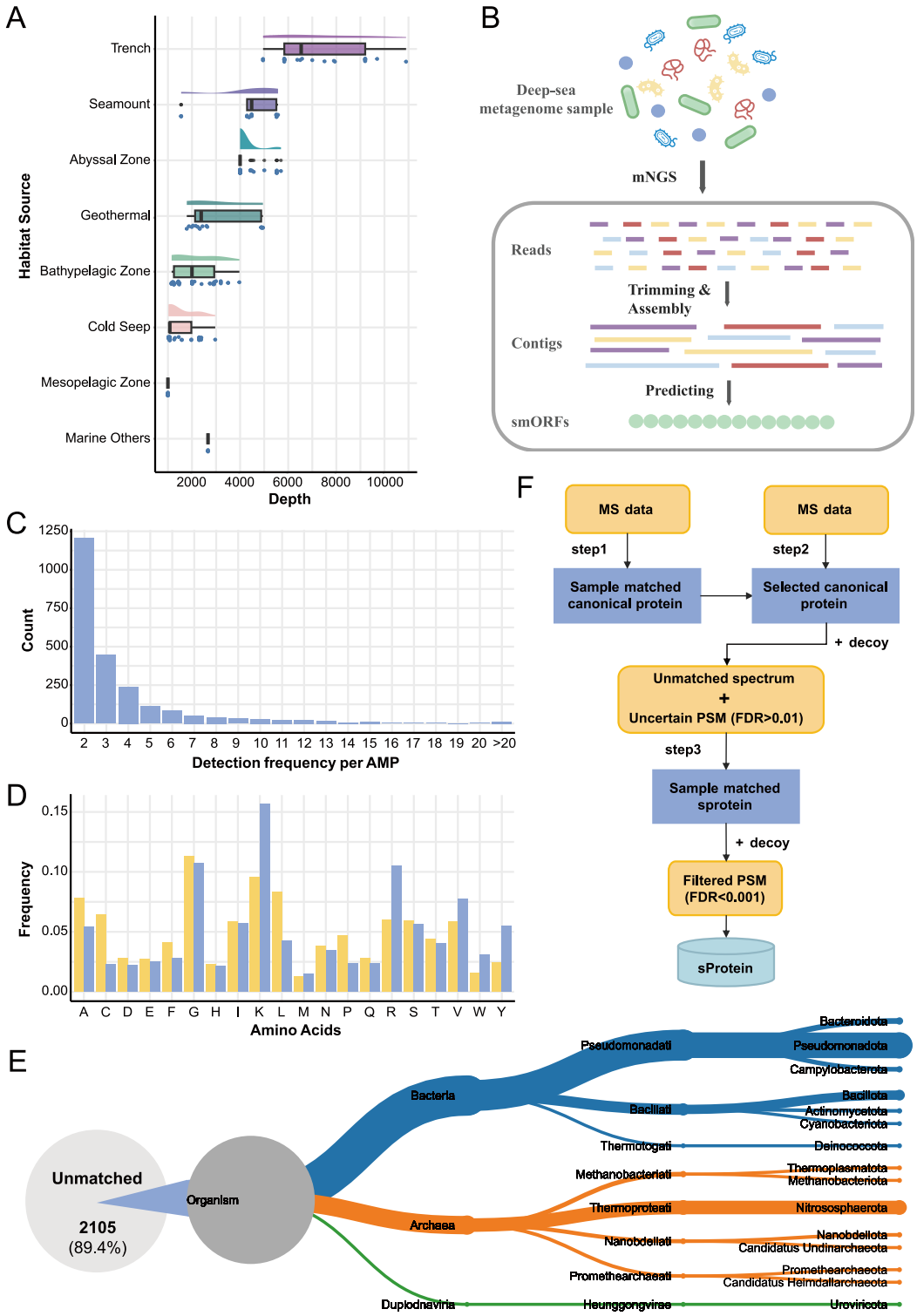


Figure 6. Systematic mining and validation of AMPs from deep-sea microorganisms.

(A). Depth Distribution of AMPs across 238 metagenomic samples from diverse marine habitats.

(B) Workflow for predicting small open reading frames (smORFs) from metagenomic assemblies.

(C). The number of times each antimicrobial peptide from the Deep-sea AMPs Database appeared in the metagenome samples.

(D). Comparison of amino acid composition between the Deep-sea AMPs Database and the public APD database shows no significant difference (P = 0.8695, Wilcoxon rank-sum test).

(E). Taxonomic lineage of the deep-sea AMPs as determined by Unipept.

(F). Three-step metaproteomic search strategy for validating the expression of predicted small proteins (sProteins) from mass spectrometry (MS) data.

A total of 31,364,546 non-redundant small proteins were predicted from 238 deep-sea metagenome samples spanning eight distinct marine habitat types (Figure 6 A, B). To enhance reliability, only sequences detected in more than one sample (2,183,935 sequences in total) were retained and subsequently screened for AMPs. Using the XAMP prediction model, a high-confidence set of 2,335 AMPs was identified, which constitute the Deep-sea AMPs Database.

Analysis of the distribution of these AMPs across metagenomic samples revealed that 1,206 were found in exactly two samples, while 12 were detected in over 20 samples, suggesting a subset possesses broad ecological prevalence (Figure 6C). Comparative analysis showed no significant difference in the overall amino acid composition between the Deep-sea AMPs Database and the public APD database (P = 0.4304, Wilcoxon rank-sum test, Figure 6D). Nevertheless, the deep-sea AMPs were enriched 1.53- to 2.93-fold in tyrosine (Y), tryptophan (W), arginine (R), and lysine (K). Taxonomic classification via Unipept[18] revealed that 89.4% of the AMPs were unassignable, indicating they likely originate from novel, unannotated smORFs (Figure 6E).

To assess expression, we analyzed metaproteomic data using a three-step database search strategy against customized databases built from sample-matched metagenomic ORFs (Figure 6F). From 2,057 identified microbial small proteins, XAMP predicted 5 to have antimicrobial activity. Notably, 4 of these overlapped with candidates in our Deep-sea AMPs Database, providing proteomic evidence for their expression (Supplementary Table 3, Figure S1).

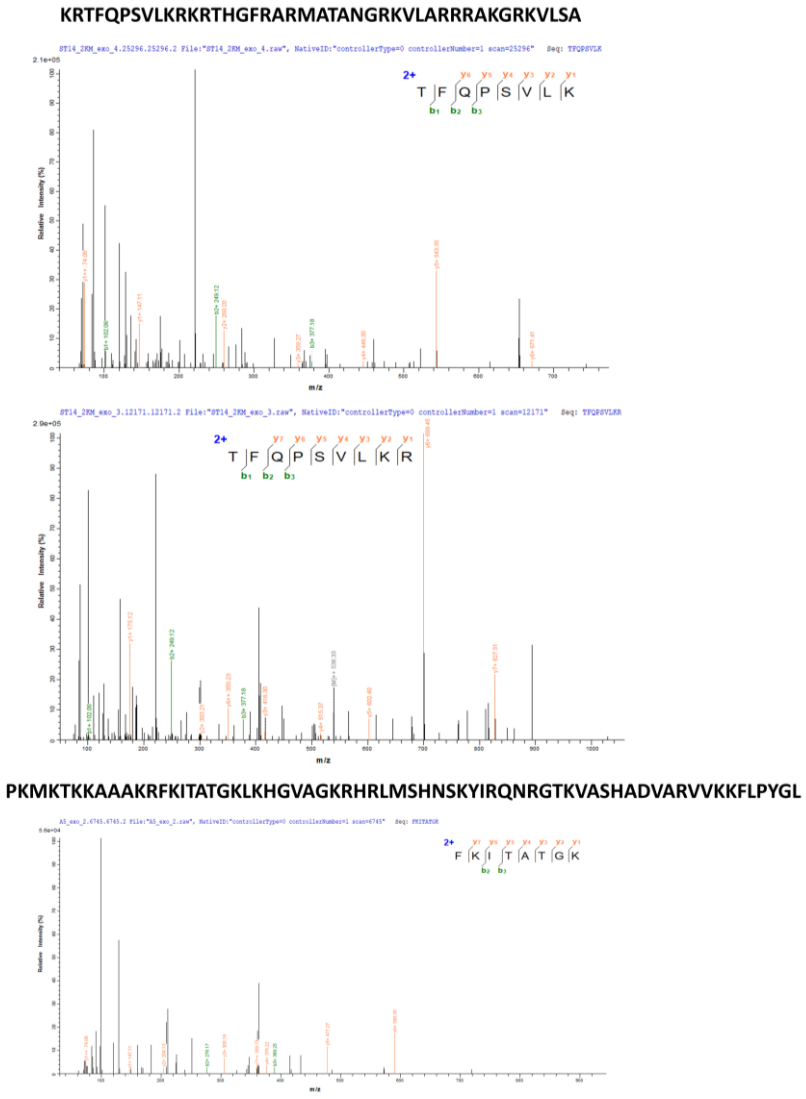
270 ***Supplementary Table 3. AMPs with spectral evidence in the metaproteome data.***

AMP sequence	AMP probability	Spectra count	Pride ID	Metagenomic evidence
PKMKTKKAAAKRFKITATGKLKHGVAGKRHRLMSHNSKYIRQNRGTKVASHADVARVVKKFLPYGL	0.597	1	PXD034421	Yes
KRTFQPSNIKRKRNHGFRARMATKNGRKIVAARRAKGRKRLTA	0.572	6	PXD034421	Yes
ARSLKKGPYIEHHLVKKVDVMNESGKKSVIKTWSRRSMISPDFVGHTFAVHNGNKFIPVFVTDNMVGHKLGEFAPTRNFRGHIAKKDKGKR	0.571	3	PXD034421	No
ATKKAGGSSRNGRDSIGRRLGVKKFGGENVLAGNIIVRQRGTKFHPGNNVGIGKDHTIFATKNGKVAFKKTRIRTFISVIPA	0.553	3	PXD034421	Yes
KRTFQPSVLKRKRTHGFRARMATANGRKVLARRRAKGRKVLSA	0.527	2	PXD034421	Yes

271

272 ***Figure S1. Example peptide-spectrum matches (PSMs) for two AMPs identified in metaproteomic***

273 ***datasets.***



274

Discussion

The global antimicrobial resistance crisis necessitates innovative therapeutic strategies. Here, we present XAMP, a dual-engine deep learning framework that overcomes critical biases in AMP prediction to enable systematic discovery of bioactive peptides from deep-sea microbiomes—an underexplored resource with high therapeutic potential.

XAMP's dual-engine design combines the high-accuracy XAMP-E and ultra-efficient XAMP-T to overcome dataset biases, yielding a superior performance gain of 0.081–0.218 in AUC. Applied to 238 deep-sea samples, it identified 12,525 high-confidence AMPs with distinct physicochemical signatures. Metaproteomic validation confirmed the expression of several predictions, with six candidates exhibiting antibacterial activity, thereby establishing a direct link from in silico discovery to experimental function.

Several limitations warrant consideration. The lower generalizability of bacteria-specific models highlights the challenge of leveraging niche biological data without sufficient sample size or specialized algorithmic design. Additionally, our analysis may underrepresent ultra-short AMPs due to prediction length cutoffs, and current validation is constrained by limited deep-sea proteomic datasets.

Future efforts should focus on expanding taxonomic diversity in training data, developing few-shot learning techniques for specialized prediction tasks, and integrating structural features to enhance prediction accuracy. Scaling metaproteomic validation and conducting systematic functional characterization will further accelerate therapeutic development.

This study establishes a robust pipeline for targeted antimicrobial discovery from extreme environments. By combining debiased deep learning with multi-omics validation, our work provides a scalable framework for addressing the urgent threat of drug-resistant pathogens through the rational exploration of microbial dark matter.

Methods

Dataset construction

The unannotated-origin training dataset was constructed by collecting antibacterial and antifungal AMPs from iAMPCN's positive dataset, which was curated from 13 public sources and 25 prior model training sets. Sequences were filtered to ≤300 amino acids (aa), yielding 13,967 peptides[8]. For the generation of negative samples at a 20:1 ratio, two steps were employed: (i) Short proteins

were extracted from UniRef90, excluding AMP-related annotations such as antimicrobial, antibacterial, antifungal, anticancer, antiviral, antiparasitic, antibiotic, antibiofilm, and effector; (ii) The length distribution of negative samples was matched to that of positive samples. Given the limitations of UniRef, sequences <10 aa were supplemented from UniProt. After the removal of duplicates, a total of 232,245 negative sequences (≤300 aa) were obtained. Notably, for all negative samples in this dataset, the methionine at the N-terminus was removed during the preprocessing stage to ensure consistency in sequence characteristics.

For the bacterial-origin training data, experimentally validated antibacterial/antifungal AMPs were sourced from APD3[19], DRAMP[20], and DBAASP[21], resulting in 598 peptides. The negative samples were composed of: (i) Short non-AMP annotated bacterial proteins from UniProt, with their lengths stratified to match those of positive samples; (ii) Supplemental sequences from the Global Microbial smORFs Catalogue (GMSC)[14] to address the scarcity in UniProt. This process generated 6,598 negative sequences (≤300 aa). During the preparation of these negative samples, the N-terminal methionine was also systematically removed. The above two datasets together constitute our mix training dataset.

An independent test set was constructed using Xiao et al.'s benchmark dataset[22]. Only antibacterial and antifungal AMPs were retained, and all sequences overlapping with our mix training dataset were removed to avoid evaluation bias. For the negative samples within this test set, the N-terminal methionine was similarly removed to maintain the same preprocessing standard as the training datasets.

Model development

Peptide representation

Initially, the character '*' was padded to the right end of each peptide sequence to achieve a maximum length of 300. Subsequently, we designed tokens for 20 common amino acids, the padding character '*' and the unknown amino acid symbol 'X'. Ultimately, each token in the peptide sequence was mapped to a distinct integer value, with 'X' corresponding to 0, 'Y' to 1, 'S' to 2, and so forth, up to '*' which was assigned the value 21.

XAMP-T architecture

The XAMP-T model is composed of an embedding layer, a positional encoder layer, a single transformer layer, and two FC layers. The processed sequences are padded/truncated to 300 amino acids using the '*' character, with a masking mechanism applied to distinguish valid sequence content from padding.

The embedding layer converts each token into a 128-dimensional word vector, with its parameters being optimized during the training phase. Sinusoidal positional encoding is employed to encode the positional information of the input sequence. The encoded sequence is subjected to a self-attention layer with eight heads that incorporates the padding mask to exclude '*' symbols from attention calculations.

Subsequently, all word vectors, excluding the padding symbol '*', are averaged to yield a 128-dimensional feature vector that characterizes the peptide sequence information. This 128-dimensional feature vector is then fed into two FC layers, the first with 1024 neurons and the second with 256 neurons. Following these two FC layers, batch normalization is applied to accelerate the convergence of the neural network. Both layers utilize the Rectified Linear Unit (ReLU) as their activation function. In the final stage, the prediction score is computed by the output of the final dense layer, which contains a single neuron. The sigmoid function serves as the activation function, confining the output to the range [0, 1]. A score of 0 indicates that the peptide is classified as non-AMP, whereas a score of 1 signifies an AMP.

XAMP-E architecture

The XAMP-E model leverages the pre-trained ESM-2 protein language model (version: *esm2_t12_35M_UR50D*)[23]. Feature representations extracted from ESM-2 are fed into a stack of FC layers for AMP/non-AMP classification. During training, the weights of the ESM-2 component are frozen, and only the parameters of the FC layers are updated.

Loss function

Given the imbalance in the datasets for antimicrobial functional activities, the focal loss function was adopted as follows:

$$FL(p_t) = -\alpha_t(1-p_t)^{\gamma}\log(p_t)$$

Where α_t is a hyperparameter that modulates the importance of samples in the positive and negative classes, while p_t represents the probability estimate produced by the model. The term $(1-p_t)^{\gamma}$ acts as a modulating factor that adjusts the impact of easily classified samples. In this implementation, the values of α_t and γ were fixed at 1 and 2, respectively.

Training process

The model development adhered to a rigorous training protocol designed to prevent data leakage and ensure robust generalization. The complete dataset was initially stratified into training set (90%) and test set (10%), preserving class distributions in both subsets. Within the training set, a 10-fold cross-validation framework was employed for simultaneous optimization of model architecture, hyperparameters, and training epochs. For each fold, training was strictly capped at 50 epochs, while

90% of the training subset was used for parameter updates while the remaining 10% served as validation monitors, with early stopping automatically triggered if validation AUC plateaued for five consecutive epochs. Using these cross-validated parameters, the model underwent final training on the entire training set. Technical execution employed the AdamW optimizer (weight decay λ=0.01, batch size 128) within PyTorch 2.3.1, with CUDA 12.1 acceleration. This multi-barrier approach effectively decoupled model development from final evaluation, maintaining the integrity of performance metrics.

Evaluation metrics

We utilized a diverse set of evaluation metrics to assess the model's performance, which include the area under the receiver operating characteristic curve (AUC), the area under the precision-recall curve (AUPRC), accuracy, precision, recall, and the false positive rate (FPR):

$$accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

$$precision = \frac{TP}{TP + FP}$$

$$recall = \frac{TP}{TP + FN}$$

$$false\ positive\ rate = \frac{FP}{TN + FP}$$

where TP represents the number of true positive instances, FP represents the number of false positive instances, TN represents the number of true negative instances, and FN represents the number of false negative instances.

Collection of deep-sea metagenome and metaproteome

In this study, we selected and downloaded 238 datasets with sampling depths greater than 1,000 meters through the sample view interface of MASH-Ocean[24] (as of June 2024), which is a platform designed for the systematic and quantitative analysis of microbial communities in marine ecosystems. Given the current challenges in protein sampling and extraction from deep-sea microbial communities, only two deep-sea metaproteomic datasets were retrieved from the PRIDE database for this study. Dataset PXD010074[25] was obtained from hydrothermal vent sites and dataset PXD009105[26] was collected from vent chimneys.

Construction of a deep-sea small protein database

We performed smORF prediction on 238 deep-sea metagenomic sequencing datasets. Raw reads

398 were filtered using Trimmomatic (v0.39), followed by assembly with MEGAHIT (v1.2.9) using the
399 parameter set "--presets meta-sensitive." Gene prediction was conducted using a modified version
400 of Prodigal with the "-p meta" parameter, targeting DNA sequences ranging from 33 to 303 bp in
401 length. Finally, redundant small protein sequences were removed to obtain a non-redundant smORF
402 dataset.

403 Deep-sea AMP identification and validation

404 The peptides in the smORF dataset were screened using the XAMP model, and candidates with
405 prediction scores >0.5 were classified as potential AMPs. To ensure their clinical viability, additional
406 selection criteria were implemented. ToxinPred[27] and HemoPI[28] were employed to exclude
407 peptides exhibiting cytotoxicity or hemolytic activity, respectively. Subsequently, the modlAMP[29]
408 Python package was used to select peptides with a net charge exceeding +2. The AxPEP[17] web
409 server (https://app.cbbio.online/ampep/home) was then used to predict the minimum inhibitory
410 concentrations (MICs) against Staphylococcus aureus and Escherichia coli. Peptides demonstrating
411 both low predicted MIC values (≤20 μg/mL) and short lengths (≤ 40 amino acids) were prioritized
412 for synthesis and validation, yielding seven final candidates.

413 Strategy for deep-sea small protein database search

414 The construction of the deep-sea metaproteomic search database followed the same workflow as
415 described for the deep-sea small protein database. In the identification of small proteins from deep-
416 sea metaproteomic data, we implemented a three-step search strategy combined with a target-decoy
417 approach to balance proteome-wide coverage and false positive control.

418 In the first step, MS-GF+ was used to search against a conventional protein database without the
419 inclusion of a decoy database. In the second step, the matched protein entried from the initial search
420 were used to construct a refined target-decoy database, and false discovery rate (FDR) control was
421 applied (FDR ≤ 1%). Peptide-spectrum matches (PSMs) passing this threshold were filtered out,
422 and a new MGF file was generated using the remaining unmatched spectra. In the third step, the
423 unmatched spectra were searched against a customized small protein database with a more stringent
424 FDR threshold of 0.1% to enhance the reliability of small protein identification. Key search
425 parameters were as follows: protease was set to Trypsin, with full enzyme cleavage specificity;
426 precursor mass tolerance was set to 10 ppm; the maximum number of modifications allowed per
427 peptide was 3; and variable modifications included methionine oxidation.

Data availability

All code used in this study and the final trained models are provided in our public GitHub repository: https://github.com/Li-Lab-SJTU/XAMP.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (grant nos.: 32570783， 32170664), the Key Project for Computational Biology of Shanghai (grant no.: 23JS1400800), and the Fundamental Research Funds for the Central Universities (grant nos.: YG2023ZD11). The computations in this article were run on the Siyuan-1 cluster supported by the Center for High Performance Computing at Shanghai Jiao Tong University. The authors thank Zhuoqi Zheng for his valuable feedback on the manuscript.

Authorship contributions

Bairun Chen: Investigation, Data Curation, Methodology, Statistical analysis, Writing - Original Draft, Visualization. **Xinyi Mou:** Investigation, Data Curation, Statistical analysis, Writing - Original Draft. **Zhuoxuan Song:** Investigation, Data Curation, Omics pipeline design and implementation. **Huaying Lin:** Methodology. **Yu Zhang:** Supervision, Project Administration, Funding Acquisition. **Jing Li:** Conceptualization, Writing - Review & Editing, Supervision, Project Administration, Funding Acquisition. All authors have read and approved the final version of the manuscript.

Declaration of competing interest

The authors declare no competing interests.

References

1. Theuretzbacher, U., et al., *Analysis of the clinical antibacterial and antituberculosis pipeline.* Lancet Infectious Diseases, 2019. **19**(2): p. E40-E50.
2. Spellberg, B., et al., *Novel Approaches Are Needed to Develop Tomorrow's Antibacterial Therapies.* American Journal of Respiratory and Critical Care Medicine, 2015. **191**(2): p. 135-140.
3. Carratalá, J.V., et al., *Nanostructured antimicrobial peptides: The last push towards clinics.* Biotechnology Advances, 2020. **44**.
4. Magana, M., et al., *The value of antimicrobial peptides in the age of resistance.* Lancet Infectious Diseases, 2020. **20**(9): p. E216-E230.
5. Zhang, Q.Y., et al., *Antimicrobial peptides: mechanism of action, activity and clinical potential.* Military Medical Research, 2021. **8**(1).
6. Santos, C.J.r., et al., *Macrel: antimicrobial peptide screening in genomes and metagenomes.* Peerj, 2020. **8**.
7. Ma, Y., et al., *Identification of antimicrobial peptides from the human gut microbiome using deep learning.* Nature Biotechnology, 2022. **40**(6): p. 921-931.
8. Xu, J., et al., *iAMPCN: a deep-learning approach for identifying antimicrobial peptides and their functional activities.* Briefings in Bioinformatics, 2023. **24**(4).
9. Cao, Q.S., et al., *Designing antimicrobial peptides using deep learning and molecular dynamic simulations.* Briefings in Bioinformatics, 2023. **24**(2).
10. Xu, J.Q., et al., *Waste to resource: Mining antimicrobial peptides in sludge from metagenomes using machine learning.* Environment International, 2024. **186**.
11. Santos, C.J.r., et al., *Discovery of antimicrobial peptides in the global microbiome with machine learning.* Cell, 2024. **187**(14).
12. Guan, C.G., et al., *Computational exploration of global venoms for antimicrobial discovery with Venomics artificial intelligence.* Nature Communications, 2025. **16**(1).
13. Torres, M.D.T., F.P. Wan, and C. de la Fuente-Nunez, *Deep learning reveals antibiotics in the archaeal proteome.* Nature Microbiology, 2025. **10**(9).
14. Duan, Y.Q., et al., *A catalog of small proteins from the global microbiome.* Nature Communications, 2024. **15**(1).
15. Giglione, C., A. Boularot, and T. Meinnel, *Protein N-terminal methionine excision.* Cellular and Molecular Life Sciences, 2004. **61**(12): p. 1455-1474.
16. Chen, N. and C. Jiang, *Antimicrobial peptides: Structure, mechanism, and modification.* European Journal of Medicinal Chemistry, 2023. **255**.
17. Cai, J.X., et al., *BERT-AmPEP60: A BERT-Based Transfer Learning Approach to Predict the Minimum Inhibitory Concentrations of Antimicrobial Peptides for*

and. Journal of Chemical Information and Modeling, 2025. **65**(7): p. 3186-3202.

18. Vande Moortele, T., et al., *Unipept in 2024: Expanding Metaproteomics Analysis with Support for Missed Cleavages and Semitryptic and Nontryptic Peptides.* Journal of Proteome Research, 2025. **24**(2): p. 949-954.
19. Wang, G.S., X. Li, and Z. Wang, *APD3: the antimicrobial peptide database as a tool for research and education.* Nucleic Acids Research, 2016. **44**(D1): p. D1087-D1093.

20. Shi, G.B., et al., *DRAMP 3.0: an enhanced comprehensive data repository of antimicrobial peptides.* Nucleic Acids Research, 2022. **50**(D1): p. D488-D496.
21. Pirtskhalava, M., et al., *DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics.* Nucleic Acids Research, 2021. **49**(D1): p. D288-D297.
22. Xiao, X., et al., *iAMP-2L: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types.* Analytical Biochemistry, 2013. **436**(2): p. 168-177.
23. Lin, Z., et al., *Language models of protein sequences at the scale of evolution enable accurate structure prediction.* bioRxiv, 2022: p. 2022.07.20.500902.
24. Wang, Y.Z., et al., *MASH-Ocean 1.0: Interactive platform for investigating microbial diversity, function, and biogeography with marine metagenomic data.* Imeta, 2024. **3**(3).
25. Meier, D.V., et al., *Microbial metal-sulfide oxidation in inactive hydrothermal vent chimneys suggested by metagenomic and metaproteomic analyses.* Environmental Microbiology, 2019. **21**(2): p. 682-701.
26. Pjevac, P., et al., *Metaproteogenomic Profiling of Microbial Communities Colonizing Actively Venting Hydrothermal Chimneys.* Frontiers in Microbiology, 2018. **9**.
27. Gupta, S., et al., *Approach for Predicting Toxicity of Peptides and Proteins.* Plos One, 2013. **8**(9).
28. Chaudhary, K., et al., *A Web Server and Mobile App for Computing Hemolytic Potency of Peptides.* Scientific Reports, 2016. **6**.
29. Müller, A.T., et al., *modlAMP: Python for antimicrobial peptides.* Bioinformatics, 2017. **33**(17): p. 2753-2755.