RESEARCH Open Access

UniAMP: enhancing AMP prediction using deep neural networks with inferred information of peptides

Zixin Chen[1], Chengming Ji[1], Wenwen Xu[1], Jianfeng Gao[4], Ji Huang[3], Huanliang Xu[1], Guoliang Qian[2]* and Junxian Huang[1]*

*Correspondence: glqian@njau.edu.cn; jim@njau.edu.cn

[1] College of Artificial Intelligence, Nanjing Agricultural University, Weigang No.1, Nanjing 210095, Jiangsu, China
[2] College of Plant Protection, Nanjing Agricultural University, Weigang No.1, Nanjing 210095, Jiangsu, China
[3] College of Agriculture, Nanjing Agricultural University, Weigang No.1, Nanjing 210095, Jiangsu, China
[4] StarHelix Inc, Jiangmiao Road, Nanjing 210000, Jiangsu, China

Abstract

Antimicrobial peptides (AMPs) have been widely recognized as a promising solution to combat antimicrobial resistance of microorganisms due to the increasing abuse of antibiotics in medicine and agriculture around the globe. In this study, we propose UniAMP, a systematic prediction framework for discovering AMPs. We observe that feature vectors used in various existing studies constructed from peptide information, such as sequence, composition, and structure, can be augmented and even replaced by information inferred by deep learning models. Specifically, we use a feature vector with 2924 values inferred by two deep learning models, UniRep and ProtT5, to demonstrate that such inferred information of peptides suffice for the task, with the help of our proposed deep neural network model composed of fully connected layers and transformer encoders for predicting the antibacterial activity of peptides. Evaluation results demonstrate superior performance of our proposed model on both balanced benchmark datasets and imbalanced test datasets compared with existing studies. Subsequently, we analyze the relations among peptide sequences, manually extracted features, and automatically inferred information by deep learning models, leading to observations that the inferred information is more comprehensive and non-redundant for the task of predicting AMPs. Moreover, this approach alleviates the impact of the scarcity of positive data and demonstrates great potential in future research and applications.

Keywords: Antimicrobial peptides, Deep learning, Feature extraction, Protein language model

Introduction

Currently, antimicrobial resistance (AMR) in bacterial infections has emerged as a critical global concern, taking precedence on the agendas of policymakers and public health authorities in both developed and developing countries [1]. For example, Gram-negative bacteria, such as CRE and members of ESKAPE (*K.pneumoniae*, *A.baumannii*, *P. aeruginosa* and *Enterobacter* spp), are of popular concern [2]. AMR transmission in agriculture involves not only foodborne pathogens but also commensals and environmental

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by-nc-nd/4.0/.

microbes, posing risks to human health from animal and plant-based foods [3]. Despite this, the reality is that investments in research and development of new antibiotics by the pharmaceutical industry and biotechnology companies are decreasing due to high failure rates and low profitability [4]. As a result, tackling AMR has posed a tremendous challenge.

Developing medicines based on antimicrobial peptides (AMPs) is a very promising solution to this global challenge. AMPs are low molecular weight proteins with broad-spectrum antimicrobial properties and immune-modulatory effects [5]. Their unique antimicrobial mechanisms reduce the likelihood of resistance development, which distinguishes them from typical antibiotics [6–8]. Therefore, AMPs serve as a promising therapeutic option, ubiquitous in the innate immune systems of various life forms [9].

Due to the time-consuming and labor-intensive nature of high-throughput experiments for evaluating each individual AMP, the accelerated advancement of research and application of AMPs is hindered. Fortunately, various databases are developed to offer information for enhancing the efficient discovery and design of AMPs. These databases empower users to explore and extract extensive details regarding peptide structures, chemical modifications, bioactivities, and classifications [10]. Most of these AMP databases contain antimicrobial targets of the AMPs and whether the AMPs are natural or synthetic. Researchers can consult these databases and obtain AMP-related information accordingly. However, the number of AMPs in each of these databases is not substantial, usually in the thousands. Additionally, the number of AMPs targeting a specific pathogen is often only in the hundreds, not to mention that there are a large number of duplicate AMP entries among different databases [11]. Compared to the vast number of peptides, the number of those with known antimicrobial activity is just a drop in the ocean, indicating that there are still many potential AMPs yet to be discovered.

With the development of artificial intelligence technologies, using computational methods to discover and design AMPs has become a trending research topic. In the last decade, several tools have been developed with Machine Learning methods: AntiCP2.0 [12] (Support Vector Machine), AmpGram [13] (Random Forest), and TP-MV [14] (ensemble ML method). In recent years, there have also been tools designed based on Deep Learning methods, a brand new and powerful branch of ML methods, including Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM): Deep-AmPEP30 [15] (CNN), sAMP-PFPDeep [16] (CNN), and AMPlify [17] (LSTM). Most of these methods directly make predictions purely based on the amino acid sequence of a candidate AMP. In addition, although iAMP-Attenpred [18] and AMP-BERT [19] also predict AMPs based on amino acid sequences, their use of BERT [20] for word semantic embedding has effectively improved accuracy and advanced the progress of AMP prediction.

Realizing that the information contained in the amino acid sequence alone might be limited, researchers also try to rely on some additional features of the peptides, such as the composition, physicochemical properties and structural properties, *etc*. The sAMP-PFPDeep [16] converts the information of the position, frequency, and 12 physicochemical properties of the peptide sequences into three-channel images as model inputs. Similarly, Deep-AmPEP30 uses the PseAAC [21] feature to predict AMPs [15]. In particular, the increasing emphasis on predicting AMPs using structural properties arises from the notably accurate predictions of protein structures by AlphaFold [22] and trRosetta [23], for example, sAMPpred-GAT uses the peptide structures predicted by trRosetta to predict the AMPs based on the Graph Attention Network (GAT) [24].

Existing studies have demonstrated the feasibility of using peptide sequence order, composition, physicochemical properties and structural properties to predict AMPs, with considerable performance. However, based on our study, the aforementioned manually extracted features for describing a candidate peptide sequence might not be sufficient for the best AMP prediction performance. We evaluated the combination of several manual feature extraction methods and found that feature concatenation may even make the feature vector less comprehensive, possibly due to more information conflict and redundancy, making model prediction more difficult. Instead, we believe that using deep learning for feature extraction might produce a better description of a peptide sequence for AMP prediction. Subsequently, we evaluated the ablation of three feature vectors extracted by deep learning models (UniRep [25], ESM-2 [26], ProtT5 [27]) and found that the combination of features from UniRep and ProtT5 was more comprehensive, leading to more accurate model predictions. In this study, we proposed an AMP prediction framework, UniAMP, which utilizes the feature information jointly inferred by the deep learning models UniRep and ProtT5. At the core of UniAMP, we designed a novel deep neural network as a predictor, composed of fully connected modules and self-attention mechanisms.

Some researchers have a need to predict AMPs targeting specific pathogens, which most previous predictors were unable to address. We selected *P. aeruginosa* as a representative bacterium and *C. albicans* as a representative fungus, which pose significant threats to plant and human health respectively and are relatively common. Additionally, we tested our model on *Salmonella spp.*, which has a smaller dataset, to further validate UniAMP's performance. In order for fair and comprehensive comparison, we aggregate benchmark datasets consisting of all the AMP entries from CAMPR4 [28], DBAASPv3 [29], dbAMP2 [30], DRAMP3 [31], LAMP2 [32] and YADAMP [33], with 9241 AMPs in total. Evaluation results on the benchmark datasets show that UniAMP clearly outperforms the existing methods under several comprehensive metrics, e.g. Matthews Correlation Coefficient(MCC) and F1-score, *etc.* Moreover, we assessed several state-of-the-art models on the test datasets, and UniAMP consistently demonstrated outstanding performance. We analyzed the inferred information and manually extracted features, concluding that the inferred information is more comprehensive and effective for AMP prediction. We believe that UniAMP may boost the research and discovery of AMPs, and we make UniAMP publicly available (https://amp.starhelix.cn).

In summary, UniAMP enhances the accuracy of AMP prediction, providing researchers with a valuable tool to advance the discovery of AMPs; moreover, it highlights the value of utilizing the rich features contained in inferred information, offering a new perspective for other predictive tasks in bioinformatics.

Materials and methods

Data collection and dataset preparation

AMP and non-AMP

We collected AMP data from six public AMP databases: CAMPR4 [28], DBAASPv3 [29], dbAMP2 [30], DRAMP3 [31], LAMP2 [32] and YADAMP [33]. These databases were merged into a single AMP database, retaining only experimentally validated, non-duplicate sequences with antimicrobial activity. The database includes peptide sequences and antimicrobial activity information, represented as Key-Value pairs of Target-Minimum Inhibitory Concentration (MIC), with all MIC units converted to μg/ml. If discrepancies in MIC records across databases were found, the largest MIC value was used. Peptides were selected for the positive dataset if they had antimicrobial activity against specific pathogens, a MIC below 100 μg/ml, and a sequence length between 6 and 50 amino acids [34, 35]. The positive datasets for *P. aeruginosa*, *C. albicans*, and *Salmonella spp.* contained 4821, 2545, and 1875 sequences, respectively.

For the negative dataset, we collected 2,835,190 non-AMP sequences from UniprotKB [36] using the search condition 'length:[6 TO 50] NOT antimicrobial NOT antibiotic NOT antiviral NOT antifungal NOT fungicide NOT secreted NOT secretory NOT excreted NOT effector NOT defensin.' Sequences with antimicrobial activity were excluded after comparison with the AMP database. All sequences contained only the 20 canonical amino acids.

Training and test datasets

In order for fair comparison and assessment of model robustness, Cluster Database at High Identity with Tolerance (CD-HIT) [37] program was adopted with a 40% sequence similarity threshold (-c 0.4, minimum value) to cluster the positive sequences, ensuring that sequences in different clusters were dissimilar [38]. In each positive dataset, all data were randomly divided into training and test datasets at a ratio of 8:2, while making sure that the split is not within any cluster, and each cluster is either in the training dataset as a whole or in the test dataset. This clustering-based splitting avoids potential bias from overlapping sequence similarities between training and test datasets, ensuring that the model is evaluated on diverse and independent data. The complete separation of clusters ensures robustness across different data distributions. Due to the imbalanced number of sequences in each cluster, random division was performed multiple times until the number of sequences in the two datasets approximate the targeted 8:2 ratio.

For the negative dataset, we used a subset of sequences due to their larger number. The negative sequences were clustered into 234,148 groups using CD-HIT, and a corresponding number of negative sequences were selected to match the positive dataset size-50 times the positive sequences for training and 100 times for testing. This imbalanced data ratio was designed to reflect the rarity of AMPs in proteins [6]. Preliminary experiments with an unoptimized BERT model confirmed that these ratios were effective for model training and evaluation, as shown in Table 1. Unfortunately, Higher ratios could not be tested due to hardware limitations. The ratio of 50:1 in the training dataset balances computational feasibility and sufficient representation of negative samples, ensuring the model learns to distinguish AMPs effectively without overfitting. For the test dataset, a higher ratio of 100:1 was used to simulate a more challenging evaluation scenario. It is important to note that the ratio of 100 is merely a hypothetical assumption to represent rarity, and the actual ratio in reality may be lower. Besides, the negative dataset was also balanced to match the peptide length distribution of the positive sequences. This helps mitigate potential bias in the test dataset.

Table 1 The impact of training dataset ratio

Negative:positive	MCC (validation)	MCC (test)
1	0.9567	0.2119
2	0.9562	0.3681
5	0.9474	0.5714
10	0.9471	0.7778
20	0.9423	0.8084
50	0.9436	0.8556

The purpose of this experiment was to determine the dataset ratio

Therefore, the hyperparameters were not optimized, and the dataset differed from that used in practice

Table 2 Datasets in this study

Datasets	Positives	Negatives
P.aeruginosa training	3828	191400
P.aeruginosa test	993	99300
C.albicans training	2036	101800
C.albicans test	509	50900
Salmonella spp. training	1490	74500
Salmonella spp. test	385	38500
P.aeruginosa benchmark	993	993

Both positive and negative sequences were filtered by CD-HIT, and the similarity between training samples and test samples is < 40%

The peptide length distributions of positive and negative sequences in the same dataset are similar

Following these procedures, we constructed 6 peptide datasets according to their different antimicrobial activities, as shown in Table 2, each AMP group containing both training and test datasets.

Benchmark datasets

We constructed a balanced benchmark dataset (Table 2) for AMPs targeting *P. aeruginosa* to test the performance compared with previous methods (most test datasets in previous methods were balanced). Specifically, all positive sequences in the test dataset of *P. aeruginosa* were selected, and the same number of negative sequences with similar length distribution were randomly selected. Since there are specific requirements of different previous studies, such as sequence length less than 30 [15] and more than 40 [24], corresponding adjustments were made according to these requirements to obtain its true performance.

Feature extraction

These studies demonstrated the low sequence homology of peptide sequences, but structural similarities in corresponding functions [24, 39, 40]. Furthermore, they used computational methods to predict the structure properties of peptides and effectively predicted the peptide's antimicrobial activity and other functions through its structural properties. This highlighted the importance of feature extraction in this task. On the other hand, we used feature vectors inferred by deep learning models trained on large-scale datasets as inputs. Since these inferred features are derived from a larger dataset, they can enhance the richness and comprehensiveness of the features extracted from smaller datasets.

In this study, we represented peptides based on their sequences, composition, physicochemical properties and inferred information. More specifically, peptides were represented in three different forms as inputs to models. Firstly, the amino acid sequences of peptides were directly used as inputs in prediction. Secondly, the composition and physicochemical properties of a peptide were represented by PCA: PseAAC [21], CT [41], and AC [42]. Thirdly, the feature vectors of the peptides were computed using UniRep [25], ESM-2 [26], and ProtT5 [27], resulting in three vectors of 1900, 1280, and 1024 dimensions, respectively. Table 3 summarizes the comparison of UniRep, ESM-2, and ProtT5.

Pseudo amino acid composition (PseAAC)

PseAAC [21] is particularly valuable for capturing information about local and global sequence patterns, which can be crucial for various tasks such as protein structure prediction, function prediction, and classification [43]. The encoding of PseAAC combines the hydrophobicity, hydrophilicity, and side-chain mass of amino acids. The PseAAC values are quantified as follows:

$$X = [x_1, x_2, \cdots, x_{20}, \cdots, x_{20+\lambda}]^T (\lambda < N) \tag{1}$$

where N represents the length of sequence, λ is the number of sequence correlation factors, and here we take it as 4. The calculation formula for each element in the vector is given by following equation:

$$x_u = \begin{cases} \dfrac{f_u}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda} \theta_j} (1 \le u \le 20) \\ \dfrac{w\theta_{u-20}}{\sum_{i=1}^{20} f_i + w \sum_{j=1}^{\lambda} \theta_j} (21 \le u \le 20 + \lambda) \end{cases} \tag{2}$$

where f_i is the normalized occurrence frequency of the 20 amino acids in sequence, w is the weight factor for the sequence order effect, we choose w=0.05. And θ_j is the jth sequence correlation factor.

Table 3 Comparison of UniRep, ESM-2 and ProtT5

Method	Model	Task	Dataset	Advantages	Limitations	Length
UniRep	mLSTM	Proteingeneration	UniRef50(24 million)	EfficiencyFast	Lackslong-range-context	1900
ESM-2	Transformer	Protein representation	UniRef50 &UniRef90 (65 million)	Strong contextual modeling	High computational cost	1280/2560
ProtT5	T5	Protein language	UniRef50 &BFD (2.1billion)	Versatile Multi-task	Slower Resource-heavy	1024×n[1]

[1] The 'n' is the length of the sequence

It is changed from 1024×n to 1024 by taking the average (in this study) or the maximum value

Conjoint triad (CT)

The CT [41] method, akin to the commonly used K-mer approach for biological sequences, categorizes amino acids into 7 classes based on their types. Subsequently, with K set to 3, resulting in a frequency space of 343 (7 × 7 × 7), amino acid sequences of length N generate $N - 2$ 3-mers. The frequencies of these 3-mers are computed and assigned to the frequency space, culminating in a 343-dimensional vector representing the peptide features.

Auto covariance descriptor (AC)

The amino acid proximity effect calculated by the AC are primarily manifested in the interactions between an amino acid and a fixed number of surrounding amino acids, showing hydrophobicity (H1), hydrophilicity (H2), net charge index (NCI), Polarity (P1), polarizability (P2), solvent-accessible surface area (SASA), and side chains (SC) [42]. Initially, for an amino acid sequence of length N, a $7 \times N$ matrix is constructed based on the aforementioned physicochemical properties. In this matrix, each element $P_{i,j}$ represents the ith property of the jth amino acid. Subsequently, normalize the matrix as follows:

$$P'_{i,j} = \frac{P_{i,j} - \overline{P}}{D_j} \tag{3}$$

where $\overline{P}$ and D_j are the mean and standard deviation of the i physicochemical property over 20 amino acids. Then calculate the AC vector based on the normalized matrix as follows:

$$AC(n,i) = \frac{\sum_{j=1}^{N-n}\left(P'_{i,j} - \frac{\sum_{j=1}^{N} P'_{i,j}}{N}\right) \times \left(P'_{i,j+n} - \frac{\sum_{i=1}^{N} P'_{i,j}}{N}\right)}{N-n} \tag{4}$$

Given the minimum length requirement of 6 for AMP sequences, $n_{\max}$ was set to 5, resulting in the representation of a peptide as a 35-dimensional (7 × 5) vector.

UniRep

UniRep[25], trained on 24 million UniRef50 [44] amino acid sequences using a 1900-hidden unit Multiplicative long-/short-term-memory (mLSTM) RNNs model capable of fully learning the rich information of natural language to generate protein sequences [45], exhibits several capabilities. It successfully learns physicochemically meaningful clusters within amino acid embeddings and proves effective in partitioning structurally similar proteins. Additionally, the model showcases its semantic richness by hierarchically clustering proteins based on expert-labeled datasets and revealing correlations between internal hidden states and protein secondary structure. Notably, UniRep's single-hidden unit positively correlates with alpha-helix annotations and negatively correlates with beta-sheet annotations, suggesting the model's ability to predict secondary structure in an unsupervised manner.

In summary, UniRep is better suited for detailed secondary structure and physicochemical insights. Based on a trained UniRep model, we converted the peptide into a 1900-dimensional vector as the input to the model.

ESM-2

ESM-2 [26] employs a transformer architecture [46] with 15 billion parameters and is trained on 138 million UniRef90 sequences and 43 million UniRef50 sequences, encompassing 65 million unique sequences.This model excels in learning evolutionary, structural, and functional features from amino acid sequences, enabling accurate prediction of secondary and tertiary protein structures. ESM-2's performance was demonstrated by metrics such as Root Mean Square Deviation (RMSD), Template Modeling Score (TM-score), and contact precision, showing strong correlations with experimental data and high accuracy, with correlation coefficients indicating strong correlations for secondary structure elements and functional site annotations. Moreover, ESM-2's feature vectors capture detailed information about protein sequences, including semantic richness, evolutionary patterns, and biophysical properties. These vectors facilitate tasks such as phylogenetic inference, structural prediction, and functional annotation.

Therefore, ESM-2 is optimal for comprehensive structural and precise functional predictions. Based on the trained ESM-2 model, we converted the peptide into a 1280-dimensional vector to serve as the input for further analysis.

ProtT5

The ProtTrans project [27] includes models like ProtBERT, ProtXLNet, and ProtT5, which are used for tasks such as protein function annotation, structure prediction, and understanding the language of proteins, with ProtT5, an auto-encoder model, specifically employed for generating vector representations of protein language models. ProtT5 was trained on a dataset of approximately 2.1 billion protein sequences, utilizing 45 million sequences from UniRef50 and 2,122 million sequences from the Big Fantastic Database (BFD [47]). By extracting information from a large number of protein sequences through self-supervised learning, ProtT5, as a pre-trained model of a protein language model, assists in capturing patterns and rules within sequences. Consequently it encompassing structural, functional, evolutionary, and contextual information of protein sequences. Moreover, as part of the protein language model, these vectors demonstrate strong potential for application in various bioinformatics tasks.

ProtT5 is particularly effective in generating protein sequence embeddings that can be used in various downstream bioinformatics tasks. Based on the trained ProtT5 model, we converted the peptide into a 1024-dimensional vector to serve as the input for further analysis.

Classification models

In this study, two types of models were used for peptide classification, distinguished primarily by the variance in their input vectors. One type used traditional Natural Language Processing (NLP) models, treating the peptide sequence as a sentence composed of words representing the 20 canonical amino acid input into the model [34]. The other type incorporated the aforementioned feature vectors as input into the model.

Sequence vector model

Previously, the Neural Network Model (NNM) based on AmpScannerV2 [38] combined with the NLP algorithm had proved effective. In particular, this method performs well on datasets containing a substantial number of negative sequences, with precision several times than that of previous approaches. As part of this analysis, three proven effective NNM [34] based on NLP algorithms were used for sequence-based AMPs prediction (Fig 1).

The first model consisted of several convolutional layers and an LSTM layer as the backbone network. The second model replaced the LSTM later in the first model with an attention (ATT) layer, while the third model was BERT model based on transformer encoders [20, 46]. Like training the NLP model, we treat the amino acid sequence as a

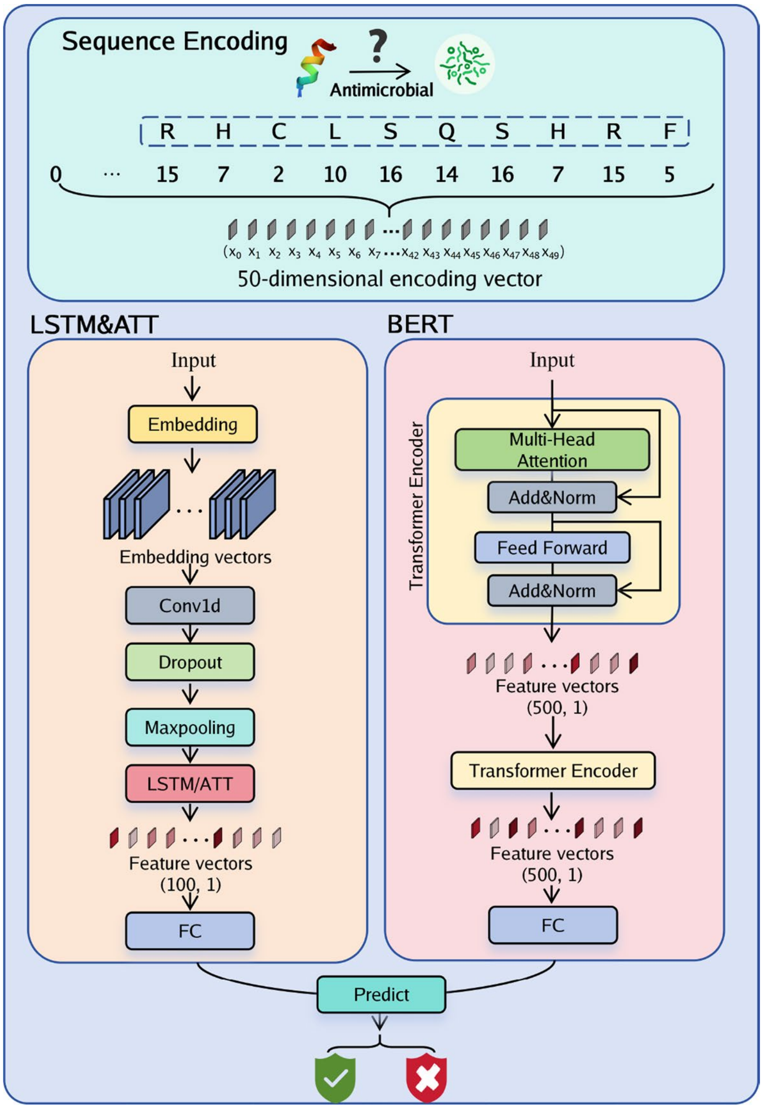

Fig. 1 The framework of Sequence Models. We used the word2vec to encode the peptide sequence into a 50-dimensional vector. Modifications to the LSTM and ATT models were limited to a single network layer, with the remaining network architecture kept consistent

sentence, with each amino acid symbol representing a word (the word vector space is of size 20 because giving the 20 canonical amino acids). Subsequently, each amino acid sequence was encoded into a 50-dimensional vector, where each dimension corresponds to the index of the amino acid symbol at that position. For sequences with fewer than 50 amino acids, zeros were padded to complete the vector.

Feature vector model

When designing the model, we started by recognizing that the feature extraction, typically handled by the backbone network, had already been completed. Consequently, our focus shifted towards constructing an effective prediction head to perform accurate classification based on the extracted features. Inspired by the study [48], we devised a network architecture named UniAMP, which incorporates two transformer encoder layers. Our objective was to facilitate the transmission of information between features and achieve feature fusion, particularly over longer distances. Unlike fully connected layers, the transformer encoders can efficiently capture relationships across distant features, thereby enriching the learned feature representations. To further enhance the model's performance and stability, we introduced batch normalization and dropout layers within the fully connected modules. Batch normalization aids in stabilizing feature mapping by normalizing intermediate outputs, while dropout helps prevent overfitting by randomly deactivating neurons during training. These additions also mitigate issues such as gradient vanishing or exploding, which are common in deep networks. In addition to the UniAMP architecture, we also implemented a baseline prediction head using a simple Multilayer Perceptron (MLP). This was done to validate our hypothesis that transformer encoders provide a significant advantage in capturing long-range dependencies and achieving better feature fusion compared to simpler models.

The model architecture is shown in Fig. 2. Initially, the feature vectors are mapped to a 256-dimensional vector through two fully connected modules. The reshaped feature vectors are then fed into two transformer encoder layers to facilitate information integration and interaction. Finally, the refined feature representations are processed through additional fully connected layers, which generate the predicted labels.

Performance measure

Evaluation metrics

Five metrics were used to evaluate the performance of different methods in this study:

$$\begin{cases} Accuracy = \dfrac{TP + TN}{TP + FP + TN + FN} \\ Precision = \dfrac{TP}{TP + FP} \\ Recall = \dfrac{TP}{TP + FN} \\ F_1 score = \dfrac{(1 + \beta^2) \times \text{Precision} \times \text{Recall}}{\beta^2 \times \text{Precision} + \text{Recall}} \\ MCC = \dfrac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \end{cases} \tag{5}$$

where *TP*, *FP*, *TN* and *FN* are the number of true positives, false positives, true negatives, and false negatives. MCC provides a balanced assessment of a model's overall

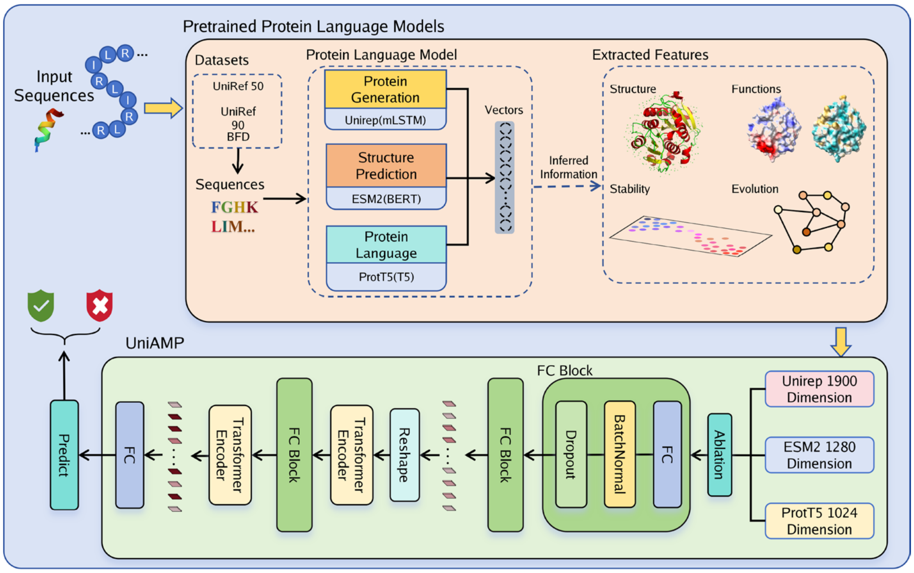

Fig. 2 The framework of UniAMP. We used trained UniRep, ESM-2, and ProtT5 as feature extractors to replace the backbone network. The inferred features from these deep learning models have proven to be information-rich. Subsequently, we input the ablation combinations of the three vectors into the model. The Transformer encoder in the model utilizes self-attention mechanisms to facilitate information transfer between features and achieve feature fusion, thereby enhancing performance

classification performance, which is especially valuable in scenarios with imbalanced class distribution [49]. In addition, Area Under the ROC Curve (AUC) and Average Precision (AP) were used, which are defined as the area under the Receiver Operating Characteristic (ROC) and Precision-Recall (PR) curve.

Model training and test

For existing AMPs predictors, their performance was evaluated on benchmark datasets using trained models published by the creators. In the case of the models in this study, we trained them on training datasets and subsequently assess their performance on both the benchmark and test datasets.

During training, we set 20% of the training data for validation. At each iteration epoch of the model, we assessed its performance on the validation set, with the MCC serving as the primary evaluation metric. It is important to note that all models shared identical training conditions, including the selection of the validation set (same random seed) and the configuration of hyperparameters (batch size=256, lr=1×10^{-4}), but we found that the models containing the transformer encoder structure is difficult to converge with a learning rate (lr) of 1×10^{-4}, so we adjust it to 1×10^{-5}. Additionally, a patience value of 30 was established, meaning that training will be halted if the model does not achieve a higher MCC within 30 epochs after reaching the current highest MCC. To enable the model to learn features of positive samples in an imbalanced dataset, we employed a criterion with weighted coefficients. This ensures that positive samples receive a higher weight when calculating the loss. We aim to eliminate potential biases in the imbalanced dataset through this approach. Besides, when training the models proposed in this paper, five models were trained strictly for each model structure using the previously mentioned training methods, and the model with the median *MCC* performance was selected for comparison.

Results

Information redundancy and overfitting of manual feature extraction methods

In this section, We evaluate various combinations of manual feature extraction methods, and their performance on *P. aeruginosa* datasets is reported in Table 4. The MCC values for three manual feature extraction methods greatly exceeded 0, confirming their informativeness. However, the combination of PseAAC, CT, and AC, which includes more feature, did not achieve the best performance, instead, PseAAC alone yielded the best performance. We observed that the combinations using PseAAC achieved similar MCC values on the validation dataset (maximum difference of 0.008), however, a notable discrepancy emerged on the test set (maximum difference of 0.043). The poorest performance observed in the combination of PseAAC and CT, particularly considering the higher dimensionality of CT compared to the other two features, led us to hypothesize that one contributing factor is the presence of additional redundant information causing the model to overfit [50]. While the model still demonstrates capability in handling higher-dimensional inputs (evidenced by similar performance on the validation dataset), in practice, it exhibits signs of overfitting. Another contributing factor is that AC and CT fail to contribute additional meaningful information compared to PseAAC.

The combination of inferred information from UniRep and ProtT5 is suitable for AMP prediction

We conducted ablation experiments to evaluate the performance of three deep learning models (UniRep, ESM-2, and ProtT5) on the downstream bioinformatics task of AMP prediction. The experiments were carried out on the *P. aeruginosa* dataset, and the results are reported in Table 5. The evaluation results for single inferred information show that, regardless of the model, the rankings for Recall, F1-score, and MCC metrics are always in the order of ProtT5, ESM-2, and UniRep from highest to lowest, while the rankings for Precision are ESM-2, UniRep, and ProtT5. It is normal for different models to have their own strengths and weaknesses, but here we mainly discuss their comprehensive performance in terms of the MCC metric. The inferred

Table 4 Ablation results of manually extracted features

Combination			Metrics	
PseAAC	CT	AC	MCC (validation)	MCC (test)
√	–	–	0.8879	**0.8315**
–	√	–	0.8357	0.7217
–	–	√	0.7720	0.6968
√	√	–	0.8864	0.7877
√	–	√	0.8856	0.8197
–	√	√	0.8802	0.7997
√	√	√	**0.8936**	0.8161

The validation set did not participate in training but was used for model selection, whereas the test set data was solely used for evaluation

Bold values indicate the maximum value

information from the ProtT5 model is undoubtedly best, which may be attributed to its use of the largest dataset (ProtT5: 2.1 billion, ESM-2: 181 million, UniRep: 24 million).

Ablation results using multiple inferred information showed consistent trends across models. the combination of UniRep and ProtT5 achieved the highest F1-score and MCC, the combination of ESM-2 and ProtT5 achieved the highest recall, while the combination of all three models achieved the highest precision. Moreover, the performance using combined inferred information is superior to that of using single inferred information. Specifically, when comparing the highest results, MCC increased by 0.008, Precision increased by 0.015, and Recall increased by 0.01. Therefore, using combined inferred information does not easily lead to information redundancy and overfitting as manually extracted information might. Considering the comprehensive situation in the task of AMP prediction, the combination of UniRep and ProtT5 is highly recommended.

Inferred information, particularly from large pre-trained models, can capture complex relationships and abstract features from protein sequences that might be overlooked by traditional feature extraction methods. By incorporating multiple inferred information sources, we enable the model to learn a richer and more comprehensive representation of the underlying biological data. This strategy effectively enhances the accuracy of AMP prediction, making it more robust and reliable. Furthermore, The use of inferred information in AMP prediction tasks has been relatively uncommon in previous studies, which primarily relied on manually curated features or word embedding vectors. UniAMP not only incorporated these inferred features but also specifically designed a framework to integrate them, enhancing the ability to capture subtle patterns in the data. This has resulted in improved prediction accuracy, as demonstrated by the ablation experiments.

Table 5 Ablation results of manually extracted features

	Combination			**Metrics**							
Model	**UniRep**	**ESM-2**	**ProtT5**	**TP**	**FP**	**TN**	**FN**	**Precision**	**Recall**	**F$_1$-score**	**MCC**
MLP	√	–	–	838	77	99223	155	0.9158	0.8439	0.8783	0.8779
MLP	–	√	–	860	69	99231	133	0.9257	0.8660	0.8948	0.8943
MLP	–	–	√	888	83	99217	105	0.9145	0.8942	0.9042	0.9033
MLP	√	√	–	896	93	99207	97	0.9059	0.9023	0.9041	0.9031
MLP	√	–	√	890	71	99229	103	0.9261	0.8962	**0.9109**	**0.9102**
MLP	–	√	√	898	87	99213	95	0.9116	**0.9043**	0.9079	0.9070
MLP	√	√	√	886	70	99230	107	**0.9267**	0.8922	0.9091	0.9084
UniAMP	√	–	–	851	63	99237	142	0.9310	0.8569	0.8925	0.8922
UniAMP	–	√	–	862	58	99242	131	0.9369	0.8680	0.9012	0.9009
UniAMP	–	–	√	893	80	99220	100	0.9177	0.8992	0.9084	0.9075
UniAMP	√	√	–	880	64	99236	113	0.9322	0.8862	0.9086	0.9080
UniAMP	√	–	√	895	65	99235	98	0.9322	0.9013	**0.9165**	**0.9158**
UniAMP	–	√	√	903	80	99220	90	0.9186	**0.9093**	0.9139	0.9131
UniAMP	√	√	√	872	44	99256	121	**0.9519**	0.8781	0.9135	0.9134

Due to the significant imbalance between positive and negative samples in the test dataset, the accuracy at the thousandth decimal place is almost identical, thus we do not compare accuracy

Bold values indicate the maximum value

UniAMP is effective on balanced datasets.

To accurately assess the performance of UniAMP, we evaluated multiple predictors on the benchmark datasets (Table 6), including CAMPR4 [28], AmPEP [15, 51], amPEPpy [52], AMPfun [53], AmpGram [13], AMPScannerV2 [38], and sAMPpred-GAT [24]. It should be noted that the positive data in the benchmark datasets have antibacterial records against *P. aeruginosa*, which is a Gram-negative bacterium. This indicates that these positive data are all AMP. On the other hand, all negative data have no information on antibacterial activity in UniProtKB and have been filtered to remove duplicates against six AMP databases. Since the experiments on antibacterial activity in UniProtKB are not exhaustive, it cannot be proven that the negative data definitively lack antibacterial activity. However, the probability of negative data being AMP is extremely low (since proteins with antibacterial activity are rare), so they may contain only a very small number of AMPs. Therefore, we can consider such benchmark datasets to be suitable for evaluating existing predictors. Additionally, a portion of the benchmark datasets was collected from CAMPR4 (it is unclear whether the data in the benchmark dataset overlaps with the training set of any predictor). As a result, the performance of existing predictors, especially CAMPR4, may be overestimated.

The performance of each predictor is presented in Table 6. UniAMP exhibits the highest accuracy, precision, F_1-score, and MCC on benchmark datasets, and the highest recall among the six models trained in this study. Interestingly, most existing predictors have a high recall rate (>0.9), whereas the model trained in this study is more inclined towards precision. The importance of these two metrics varies depending on the scenario, making it difficult to determine which is more critical. However,

Table 6 Performance of UniAMP and some existing AMPs predictors on benchmark datasets

Mehod	TP	FP	TN	FN	Accuracy	Precision	Recall	F_1-score	MCC
CAMPR4-RF [28]	970	167	826	23	0.9043	0.8531	0.9768	0.9108	0.8173
CAMPR4-SVM [28]	935	142	851	58	0.8993	0.8682	0.9416	0.9034	0.8015
RF-AmPEP30[1] [15]	851	137	667	43	0.894	0.8613	0.9519	0.9043	0.7911
AMPScannerV2 [38]	917	172	821	56	0.8852	0.8449	0.9436	0.8915	0.7757
sAMPpred-GAT[1] [24]	48	19	94	0	0.882	0.7164	**1.0**	0.8348	0.7720
amPEPpy [52]	927	172	821	66	0.8802	0.8435	0.9335	0.8862	0.7647
AmpGram[1] [13]	808	202	660	54	0.8515	0.8000	0.9374	0.8633	0.7136
CAMPR4-ANN [28]	837	148	845	156	0.8469	0.8497	0.8463	0.8484	0.6939
AMPfun [53]	940	366	627	53	0.789	0.7198	0.9466	0.8178	0.6091
AmPEP [51]	544	418	575	449	0.5634	0.5655	0.5478	0.5565	0.1270
LSTM	847	5	988	146	0.9239	0.9941	0.8529	0.9181	0.8566
ATT	826	5	988	167	0.9133	0.9939	0.8318	0.9057	0.8379
BERT	848	4	989	145	0.925	0.9953	0.854	0.9193	0.8586
MLP	890	3	990	103	0.9466	0.9966	0.8962	0.9437	0.8978
PCA	771	6	987	222	0.8852	0.9923	0.7764	0.8712	0.7893
UniAMP	895	1	992	98	**0.9501**	**0.9988**	0.9013	**0.9475**	**0.9046**

Existing AMPs Predictors used trained models published by the creators

[1] Some predictors exhibited sample deficiencies due to their constraints, and we selected the subset meeting the constraints

Bold values indicate the maximum value

UniAMP maintains a high precision while also achieving a recall above 0.9, making it potentially suitable for a wider range of scenarios.

Subsequently, we primarily discuss MCC. The two models using inferred information, UniAMP and MLP, ranked first and second. However, when the inferred information was replaced with manually extracted information using the same model structures, their comprehensive ability significantly decreased, ranking only at a medium level among existing tools and the lowest among the six models tested in this study. Moreover, models using inferred information improved the MCC by at least 0.04 compared to models using sequence information. These results demonstrate that using inferred information, particularly UniAMP, is highly effective for this task.

It was observed that the predictors trained in this study (except for PCA) outperformed existing predictors, which may raise concerns about whether these results reflect the true performance of each predictor. As previously stated, the benchmark datasets are suitable for all predictors, and we made our best effort to avoid similarities between the test and the training datasets (length confounding and similarity filtering [37]). Based on the results, UniAMP outperformed the out-of-the-box predictors by at least 0.087 and the three proven baseline models (LSTM, ATT and BERT [34]) by 0.046. Since the three baseline models were trained on the same dataset, we believe these results demonstrate its effectiveness.

UniAMP is effective on imbalanced datasets.

The imbalanced test dataset better reflects the actual situation, where AMPs are rare. We set the ratio of positive to negative data at 1 to 100, but the real ratio may be even lower, making the model's performance on imbalanced datasets more important. The performance of all models is reported in the Table 7. In this study, we built three test datasets, differentiated by the antimicrobial activity of the positive data, which are targeted against *P. aeruginosa*, *C. albicans*, and *Salmonella spp.*. Evaluation results show that UniAMP achieved the highest scores across all metrics on the *P. aeruginosa* and *C. albicans* datasets, and on the *Salmonella spp.* dataset, it was the highest in all metrics except for Precision, where MLP scored higher.

Firstly, In terms of feature information, similar to the phenomenon shown in Table 6, models using inferred information had higher MCC scores on all three test datasets compared to sequence models. In particular, UniAMP outperformed the three sequence models in all metrics except for one precision. Secondly, In terms of effectiveness of the models. As shown in Table 5, in all combinations of the ablation experiments, UniAMP's MCC is at least 0.004 higher than that of MLP. On the three test datasets, this difference reached to 0.023. Finally, in terms of data volume, the amount of data in the three test datasets gradually decreases, which mirrors real-world conditions. *P. aeruginosa*, being a frequently studied object, has a larger amount of valid data, whereas for rarer pathogens, the data significantly decreases. UniAMP demonstrates optimal comprehensive performance across all three datasets, highlighting its applicability. In conclusion, UniAMP is highly recommended for solving AMP prediction problems. This is not only due to the comprehensiveness of the inferred information from the deep learning models but also due to the specialized network architecture built for this information.

Table 7 erformance of UniAMP and baseline models on test datasets

Dataset	Model	TP	FP	TN	FN	Precision	Recall	F_1-score	MCC
P. aeruginosa	LSTM	847	94	99206	146	0.9001	0.8529	0.8759	0.8750
P. aeruginosa	ATT	826	77	99223	167	0.9147	0.8318	0.8713	0.8710
P. aeruginosa	BERT	848	94	99206	145	0.9002	0.8539	0.8764	0.8755
P. aeruginosa	MLP	890	71	99229	103	0.9261	0.8962	0.9109	0.9102
P. aeruginosa	UniAMP	895	65	99235	98	**0.9322**	**0.9013**	**0.9165**	**0.9158**
C. albicans	LSTM	393	46	50854	116	0.8952	0.7721	0.8291	0.8298
C. albicans	ATT	398	56	50844	111	0.8767	0.7819	0.8266	0.8262
C. albicans	BERT	409	59	50841	100	0.8739	0.8035	0.8372	0.8364
C. albicans	MLP	435	70	50830	74	0.8613	**0.8546**	0.8579	0.8565
C. albicans	UniAMP	435	44	50856	74	**0.9081**	**0.8546**	**0.8805**	**0.8798**
Salmonella spp.	LSTM	297	33	38467	88	0.9000	0.7714	0.8307	0.8317
Salmonella spp.	ATT	292	34	38466	93	0.8957	0.7584	0.8213	0.8226
Salmonella spp.	BERT	293	33	38467	92	0.8987	0.7610	0.8241	0.8254
Salmonella spp.	MLP	298	30	38470	87	**0.9085**	0.7740	0.8359	0.8371
Salmonella spp.	UniAMP	313	37	38463	72	0.8942	**0.8129**	**0.8517**	**0.8512**

Due to the significant imbalance between positive and negative samples in the test dataset, the accuracy at the thousandth decimal place is almost identical, thus we do not compare accuracy

Bold values indicate the maximum value

Comprehensive performance of UniAMP in AMP prediction

In the previous evaluation, we used a confidence threshold of 0.5 for classification. To comprehensively assess the model's performance at various thresholds, we also used AUC and AP for evaluation. AUC measures a model's ability to distinguish between positive and negative classes at various thresholds, making it a crucial metric for comprehensive model evaluation compared to fixed-threshold metrics. However, on imbalanced datasets, particularly those with significant imbalance, AUC may achieve a high score due to low confidence values of a large number of negative samples, potentially overestimating the model's capability. In contrast, AP is more suitable for evaluating imbalanced datasets because it primarily captures the model's ability to maintain high precision and recall. Therefore, we use AUC to evaluate the benchmark dataset and AP to evaluate the test dataset.

The ROC and PR curves are reported in Fig. 3. On the benchmark dataset, UniAMP's AUC is only lower than that of the predictor sAMPpred-GAT [24]. However, due to the limitations of sAMPpred-GAT (sequence length greater than 40), the available samples constitute only one-tenth of the dataset, which severely distorts the evaluation results. Among the remaining predictors, UniAMP has the highest AUC, indicating that its performance is not sensitive to threshold selection. It demonstrates strong ability to distinguish between positive and negative classes across various thresholds, reflecting its excellent robustness and comprehensive performance. Additionally, UniAMP achieved the highest AP in two of the three test datasets, and ranked second in the remaining one, just behind MLP. The PR curves on the *P. aeruginosa* and *Salmonella spp.* datasets are not similar. Except for UniAMP, which is the highest, the other four models do not show any regularity. This indicates that model performance varies with different datasets, with UniAMP being relatively the most stable. Unfortunately, UniAMP did not maintain the best performance on the *C. albicans* dataset. However, it exceeded the performance of

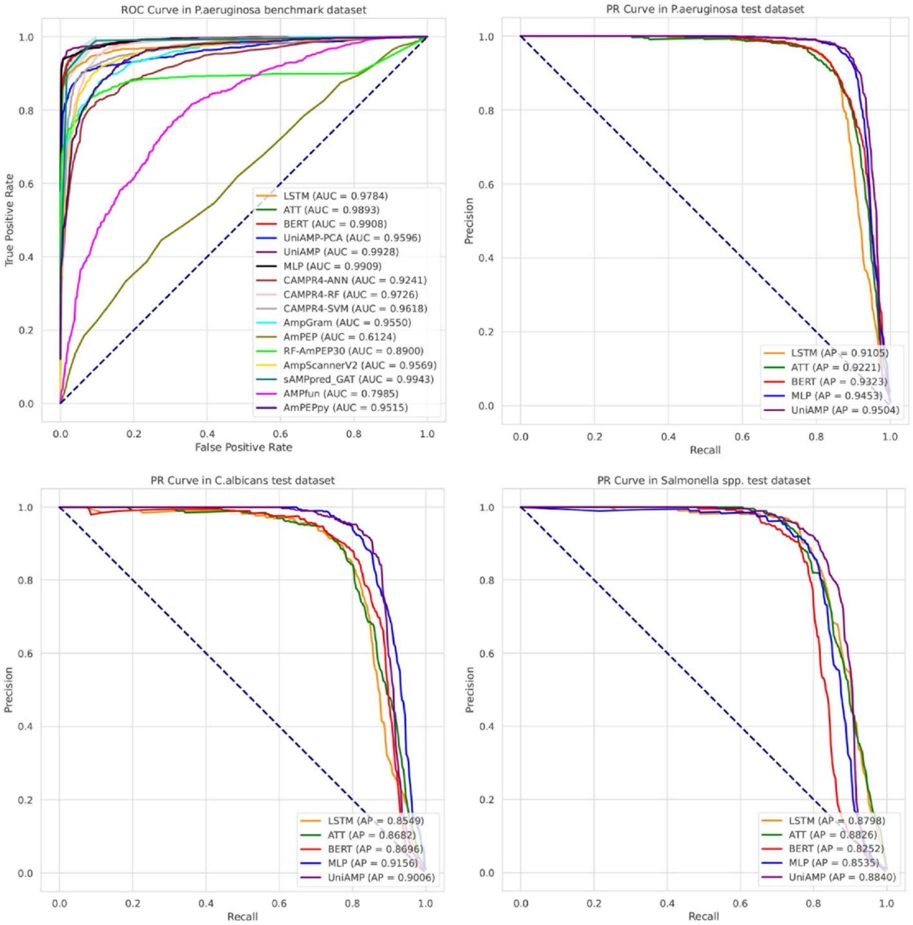

Fig. 3 AUC measures a model's ability to distinguish between positive and negative classes at various thresholds, while AP primarily captures the model's ability to maintain high precision and high recall. Therefore, AUC is more suitable for balanced benchmark datasets, whereas AP is appropriate for severely imbalanced test datasets

the three sequence models by at least 0.03. Considering its stability, the overall performance of UniAMP is commendable.

Discussion

Previously, most predictors were designed for all AMPs, providing them with a larger dataset for training. Moreover, when targeting specific pathogens, such as *E.coli* [35], the available positive data is extremely scarce, even less than one-tenth of the original dataset, making model training more challenging. As is well known, the backbone of a deep learning model is significant, as the features it extracts are directly related to the model's performance. However, relying solely on limited data makes it difficult to extract comprehensive and rich features. Many previous protein engineering models have been trained on large protein datasets. Although their tasks were not AMP prediction, the rich protein-related information learned by these models can be used for downstream bioinformatics tasks. Therefore, we believe that the inferred information from pre-trained models can replace the information from backbone trained with few-shot learning. As demonstrated in this study, UniAMP exhibits excellent performance in predicting AMPs with antibacterial activity against *P. aeruginosa*,*C. albicans* and *Salmonella spp.*. These results confirm our hypothesis. On the other hand, manual feature extraction can similarly use abundant prior knowledge to make up for data scarcity. We do not doubt that in the future, it will be possible to manually extract all functionally relevant information for AMPs comprehensively. However, manual feature extraction relies on extensive real experimental data and thorough mechanistic explanations. To the best of our knowledge, no study has systematically extracted all features relevant to the functionality of AMPs. The manually extracted information in this study performed significantly worse than the inferred information. Therefore, considering the current state of research, inferred information appears to be a better choice than manually extracted information.

Although UniAMP exhibits excellent performance, it was observed that its recall on the benchmark dataset is lower than that of several existing predictors. One possible explanation is that existing predictors are designed for all AMPs and can predict antibacterial activity against different pathogens, which may lead to some false recalls: the predicted antibacterial activity may not be specific to *P. aeruginosa*. Specifically, the positive data in the benchmark dataset not only exhibit antibacterial activity against *P. aeruginosa* but may also have antibacterial activity against other pathogens, and this likelihood is not low [54]. The existing predictors actually predict the antibacterial activity of this data against other pathogens rather than *P. aeruginosa*. This indicates the necessity of developing out-of-the-box models specifically targeting particular pathogens. Another possible explanation is that the antimicrobial mechanisms of AMPs are similar for different pathogens. This allows these models to learn additional antimicrobial mechanisms from more data to predict antibacterial activity. However, some antimicrobial mechanisms may not apply to *P. aeruginosa*, resulting in a significant decrease in precision. This is merely a hypothesis, and we hope to explore it further in future research.

Table 5 shows that various combinations of inferred information outperform single inferred information, and we recommend the combination of UniRep and ProtT5 for better overall performance. However, in different scenarios, other combinations are also worth considering. The combination of ESM-2 and ProtT5 can achieve higher recall, while the combination of all three can achieve higher precision, which should not be overlooked. This study only explored three models: UniRep, ESM-2, and ProtT5. In fact, there are many more protein engineering models available. From a combination perspective, there are numerous ablation experiments that can be conducted, making such exploration seemingly endless. However, the results show that the performance of a single ProtT5 is similar to the combination of UniRep and ESM-2. This suggests another idea: it is possible that the information inferred from a single model is comparable to that inferred from combinations. Based on the results of this study, for a single model to replace the combination of multiple models, the protein engineering model must be comprehensive. This means it requires support from a large dataset and training on multiple tasks to ensure that the information it learns is sufficiently rich. We will continue to pay attention to relevant information and use various upstream deep learning models to solve such downstream tasks.

Some researchers aim to discover specific AMPs. While many excellent predictors can predict AMPs, they lack the capability to identify which pathogens these peptides specifically target. Despite the broad-spectrum nature of AMPs, extensive trial and error is still required. Currently, researchers seeking to discover novel AMPs against pathogens like *P. aeruginosa* can leverage UniAMP for high-throughput screening of potential peptide sequences. These sequences can be obtained through various methods, such as extracting them from the proteomes of organisms known to exhibit resistance against *P. aeruginosa*. Experimentally testing sequences with the highest model-predicted scores significantly reduces redundant experiments, saving both time and cost. This framework can be adapted to other pathogens with our publicly available training method. By utilizing inferred information, UniAMP provides a more comprehensive understanding of the intrinsic mechanisms, enhancing its practical value. However, its limitations should not be overlooked. Evaluation results reveal that MCC decreases as data volume decreases, indicating that UniAMP cannot fully resolve the issue of data scarcity. Imbalanced datasets may introduce additional potential biases, including not only the model's tendency caused by the imbalance in the training data but also the uncertainty in the test dataset, which may fail to reflect the true sample space. Additionally, given the diverse antimicrobial mechanisms of AMPs, the pathogens studied here can not fully represent the entire range of pathogen populations. Consequently, UniAMP's generalizability to other pathogens and diseases requires further exploration.

In terms of the model, we have some follow-up improvement plans and suggestions. First, UniAMP was designed with a structure that includes the functions of the head and neck, while replacing the backbone with upstream deep learning models. However, during training for downstream tasks, we did not train the parameters of these upstream models. While it is possible to freeze the backbone parameters to ensure the stability of extracted features and reduce overfitting [55], this approach is still worth exploring. Secondly, although the inferred information obtained from upstream models has proven to be relatively comprehensive, whether it can further extract deep features remains to be discussed. In the future, we plan to add a deep feature extraction module to UniAMP. Third, the total set of AMPs can be viewed as the union of specific AMPs sets. Similarly, the task of predicting AMPs is the union of tasks predicting specific AMPs. This suggests that we can use a multi-task learning framework. Specifically, in this study, we built three separate models for three specific AMPs, while many predictors built one general prediction model for all AMPs. Combining both approaches, using a multi-task learning framework to build only one model that simultaneously predicts the antibacterial activity against multiple pathogens might be a good idea. The antibacterial mechanisms against different pathogens have both commonalities and differences, putting the prediction tasks of AMPs for various pathogens in a cooperative state. This might improve the model's performance on each individual task [56].

In summary, we hope this study not only provides a useful AMPs predictor, UniAMP, which uses inferred information from UniRep and ProtT5, but also offers valuable references and suggestions for upstream and downstream research in bioinformatics.

Conclusion

In this study, we proposes a framework for predicting AMPs, called UniAMP to accelerate the discovery of AMPs. The framework uses a feature vector with 2924 values inferred from two protein engineering models, UniRep and ProtT5, to represent proteins. Furthermore, we designed an advanced deep learning model for this vector to predict whether it has antimicrobial activity against specific pathogens. Evaluation results show that the performance of the model exceeds multiple existing predictors and baseline models on four evaluation datasets. We conducted multiple ablation and comparison experiments between peptide sequences, artificial information, and inferred information. The results indicate that, at this stage, the information inferred using deep learning models is more comprehensive and non-redundant. This characteristic contributes to UniAMP's excellent performance and robustness, and this framework exhibits potential applications in future research. To assist researchers with downstream tasks, we have made the data and code publicly available and released an online tool for UniAMP. Additionally, we analyzed the strengths and weaknesses of UniAMP and proposed several improvement plans and suggestions for this task, with the hope that they will be helpful to researchers in this field.

In summary, UniAMP enhances the accuracy of AMP prediction, providing researchers with a valuable tool to advance the discovery of AMPs; moreover, it highlights the value of utilizing the rich features contained in inferred information, offering a new perspective for other predictive tasks in bioinformatics.

Author contributions

ZC: Data collection and analysis, methodology, programming, experiment conducting, writing-original draft; CJ: Methodology, visualization; WX: Methodology; JG: Software; HX: Supervision, project administration; Ji H: Data analysis, writing-reviewing and editing; GQ: Conceptualization, validation and supervision, writing-reviewing and editing; Junxian H: Conceptualization, methodology, formal analysis, supervision, writing-reviewing and editing; All authors read and approved the final manuscript.

Funding

No funding.

Availability of data and materials

The data is available on https://github.com/quietbamboo/UniAMP. The complete dataset of non-AMP data can be downloaded from UniprotKB based on the conditions ('length:[6 TO 50] NOT antimicrobial NOT antibiotic NOT antiviral NOT antifungal NOT fungicide NOT secreted NOT secretory NOT excreted NOT effector NOT defensin'). All relevant links for AMPs are provided as follows:CAMPR4, DBAASPv3, dbAMP2, DRAMP3, LAMP2, and YADAMP.

Code availability

UniAMP is available on https://github.com/quietbamboo/UniAMP.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare that they have no conflict of interest.

Received: 21 August 2024 Accepted: 2 January 2025
Published online: 11 January 2025

References

1. Petrosillo N. Infections: the emergency of the new millennium. Nucl Med Infect Dis. 2020:1–8.
2. Organization WH. 2019 Antibacterial agents in clinical development: an analysis of the antibacterial clinical development pipeline. World Health Organization. 2020.
3. Thanner S, Drissner D, Walsh F. Antimicrobial resistance in agriculture. MBio. 2016;7(2):10–1128.
4. Årdal C, Balasegaram M, Laxminarayan R, McAdams D, Outterson K, Rex JH, Sumpradit N. Antibiotic development–economic, regulatory and societal challenges. Nat Rev Microbiol. 2020;18(5):267–74.

5. Zhang Q-Y, Yan Z-B, Meng Y-M, Hong X-Y, Shao G, Ma J-J, Cheng X-R, Liu J, Kang J, Fu C-Y. Antimicrobial peptides: mechanism of action, activity and clinical potential. Mil Med Res. 2021;8:1–25.
6. Boparai JK, Sharma PK. Mini review on antimicrobial peptides, sources, mechanism and recent applications. Protein Pept Lett. 2020;27(1):4–16.
7. Le C-F, Fang C-M, Sekaran SD. Intracellular targeting mechanisms by antimicrobial peptides. Antimicrob Agent Chemother. 2017;61(4):10–1128.
8. Kumar P, Kizhakkedathu JN, Straus SK. Antimicrobial peptides: diversity, mechanism of action and strategies to improve the activity and biocompatibility in vivo. Biomolecules. 2018;8(1):4.
9. Wang P, Ge R, Liu L, Xiao X, Li Y, Cai Y. Multi-label learning for predicting the activities of antimicrobial peptides. Sci Rep. 2017;7(1):2202.
10. Ramazi S, Mohammadi N, Allahverdi A, Khalili E, Abdolmaleki P. A review on antimicrobial peptides databases and the computational tools. Database. 2022;2022:011.
11. Porto W, Pires A, Franco O. Computational tools for exploring sequence databases as a resource for antimicrobial peptides. Biotechnol Adv. 2017;35(3):337–49.
12. Agrawal P, Bhagat D, Mahalwal M, Sharma N, Raghava GP. Anticp 2.0: an updated model for predicting anticancer peptides. Brief Bioinform. 2021;22(3):153.
13. Burdukiewicz M, Sidorczuk K, Rafacz D, Pietluch F, Chilimoniuk J, Rödiger S, Gagat P. Proteomic screening for prediction and design of antimicrobial peptides with ampgram. Int J Mol Sci. 2020;21(12):4310.
14. Yan K, Lv H, Wen J, Guo Y, Liu B. Tp-mv: therapeutic peptides prediction by multi-view learning. Curr Bioinform. 2022;17(2):174–83.
15. Yan J, Bhadra P, Li A, Sethiya P, Qin L, Tai HK, Wong KH, Siu SW. Deep-ampep30: improve short antimicrobial peptides prediction with deep learning. Mol Ther Nucleic Acids. 2020;20:882–94.
16. Hussain W. samp-pfpdeep: improving accuracy of short antimicrobial peptides prediction using three different sequence encodings and deep neural networks. Brief Bioinform. 2022;23(1):487.
17. Li C, Sutherland D, Hammond SA, Yang C, Taho F, Bergman L, Houston S, Warren RL, Wong T, Hoang LM, et al. Amplify: attentive deep learning model for discovery of novel antimicrobial peptides effective against who priority pathogens. BMC Genomics. 2022;23(1):77.
18. Xing W, Zhang J, Li C, Huo Y, Dong G. iamp-attenpred: a novel antimicrobial peptide predictor based on bert feature extraction method and CNN-bilstm-attention combination model. Brief Bioinform. 2024;25(1):443.
19. Lee H, Lee S, Lee I, Nam H. Amp-bert: prediction of antimicrobial peptide function based on a Bert model. Protein Sci. 2023;32(1):4529.
20. Bert DJ. Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805. 2018.
21. Chou K-C. Prediction of protein cellular attributes using pseudo-amino acid composition. Proteins: Struct Funct Bioinform. 2001;43(3):246–55.
22. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. Highly accurate protein structure prediction with alphafold. Nature. 2021;596(7873):583–9.
23. Du Z, Su H, Wang W, Ye L, Wei H, Peng Z, Anishchenko I, Baker D, Yang J. The trrosetta server for fast and accurate protein structure prediction. Nat Protoc. 2021;16(12):5634–51.
24. Yan K, Lv H, Guo Y, Peng W, Liu B. Samppred-gat: prediction of antimicrobial peptide by graph attention network and predicted peptide structure. Bioinformatics. 2023;39(1):715.
25. Alley EC, Khimulya G, Biswas S, AlQuraishi M, Church GM. Unified rational protein engineering with sequence-based deep representation learning. Nat Methods. 2019;16(12):1315–22.
26. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023;379(6637):1123–30.
27. Elnaggar A, Heinzinger M, Dallago C, Rehawi G, Wang Y, Jones L, Gibbs T, Feher T, Angerer C, Steinegger M, et al. Prottrans: toward understanding the language of life through self-supervised learning. IEEE Trans Pattern Anal Mach Intell. 2021;44(10):7112–27.
28. Gawde U, Chakraborty S, Waghu FH, Barai RS, Khanderkar A, Indraguru R, Shirsat T, Idicula-Thomas S. Campr4: a database of natural and synthetic antimicrobial peptides. Nucleic Acids Res. 2023;51(D1):377–83.
29. Pirtskhalava M, Amstrong AA, Grigolava M, Chubinidze M, Alimbarashvili E, Vishnepolsky B, Gabrielian A, Rosenthal A, Hurt DE, Tartakovsky M. Dbaasp v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. Nucleic Acids Res. 2021;49(D1):288–97.
30. Jhong J-H, Yao L, Pang Y, Li Z, Chung C-R, Wang R, Li S, Li W, Luo M, Ma R, et al. Dbamp 2.0: updated resource for antimicrobial peptides with an enhanced scanning method for genomic and proteomic data. Nucleic Acids Res. 2022;50(D1):460–70.
31. Shi G, Kang X, Dong F, Liu Y, Zhu N, Hu Y, Xu H, Lao X, Zheng H. Dramp 3.0: an enhanced comprehensive data repository of antimicrobial peptides. Nucleic Acids Res. 2022;50(D1):488–96.
32. Ye G, Wu H, Huang J, Wang W, Ge K, Li G, Zhong J, Huang Q. Lamp2: a major update of the database linking antimicrobial peptides. Database. 2020;2020:061.
33. Piotto SP, Sessa L, Concilio S, Iannelli P. Yadamp: yet another database of antimicrobial peptides. Int J Antimicrob Agents. 2012;39(4):346–51.
34. Ma Y, Guo Z, Xia B, Zhang Y, Liu X, Yu Y, Tang N, Tong X, Wang M, Ye X, et al. Identification of antimicrobial peptides from the human gut microbiome using deep learning. Nat Biotechnol. 2022;40(6):921–31.
35. Wang C, Garlick S, Zloh M. Deep learning for novel antimicrobial peptide design. Biomolecules. 2021;11(3):471.
36. Consortium U. Uniprot: a worldwide hub of protein knowledge. Nucleic Acids Res. 2019;47:506–15.
37. Fu L, Niu B, Zhu Z, Wu S, Li W. Cd-hit: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150–2.
38. Veltri D, Kamath U, Shehu A. Deep learning improves antimicrobial peptide recognition. Bioinformatics. 2018;34(16):2740–7.

39. Wei L, Ye X, Xue Y, Sakurai T, Wei L. Atse: a peptide toxicity predictor by exploiting structural and evolutionary information based on graph neural network and attention mechanism. Brief Bioinform. 2021;22(5):041.
40. Wei L, He W, Malik A, Su R, Cui L, Manavalan B. Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework. Brief Bioinform. 2021;22(4):275.
41. Shen J, Zhang J, Luo X, Zhu W, Yu K, Chen K, Li Y, Jiang H. Predicting protein-protein interactions based only on sequences information. Proceed Natl Acad Sci. 2007;104(11):4337–41.
42. Zhang L, Yu G, Xia D, Wang J. Protein-protein interactions prediction based on ensemble deep neural networks. Neurocomputing. 2019;324:10–9.
43. Chen C, Zhang Q, Ma Q, Yu B. Lightgbm-ppi: predicting protein-protein interactions through lightgbm with multi-information fusion. Chemom Intell Labor Syst. 2019;191:54–64.
44. Suzek, B.E., Wang, Y., Huang, H., McGarvey, P.B., Wu, C.H., Consortium, U. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. Bioinformatics. 2015;31(6):926–32.
45. Radford A, Jozefowicz R, Sutskever I. Learning to generate reviews and discovering sentiment. arXiv preprint arXiv:1704.01444. 2017.
46. Vaswani, A. Attention is all you need. arXiv preprint arXiv:1706.03762. 2017.
47. Steinegger M, Mirdita M, Söding J. Protein-level assembly increases protein sequence recovery from metagenomic samples manyfold. Nat Methods. 2019;16(7):603–6.
48. Li X, Han P, Wang G, Chen W, Wang S, Song T. Sdnn-ppi: self-attention with deep neural network effect on protein-protein interaction prediction. BMC Genomics. 2022;23(1):474.
49. Chicco D, Jurman G. The advantages of the Matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. BMC Genomics. 2020;21:1–13.
50. Liu R, Gillies DF. Overfitting in linear feature extraction for classification of high-dimensional image data. Pattern Recogn. 2016;53:73–86.
51. Bhadra P, Yan J, Li J, Fong S, Siu SW. Ampep: sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. Sci Rep. 2018;8(1):1697.
52. Lawrence TJ, Carper DL, Spangler MK, Carrell AA, Rush TA, Minter SJ, Weston DJ, Labbé JL. ampeppy 1.0: a portable and accurate antimicrobial peptide prediction tool. Bioinformatics. 2021;37(14):2058–60.
53. Chung C-R, Kuo T-R, Wu L-C, Lee T-Y, Horng J-T. Characterization and identification of antimicrobial peptides with different functional activities. Brief Bioinform. 2020;21(3):1098–114.
54. Koo HB, Seo J. Antimicrobial peptides under clinical investigation. Pept Sci. 2019;111(5):24122.
55. Wimmer P, Mehnert J, Condurache AP. Dimensionality reduced training by pruning and freezing parts of a deep neural network: a survey. Artif Intell Rev. 2023;56(12):14257–95.
56. Standley T, Zamir A, Chen D, Guibas L, Malik J, Savarese S. Which tasks should be learned together in multi-task learning? In: International Conference on Machine Learning. 2020; pp. 9120–9132 . PMLR

Publisher's Note

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