Journal Title Here, 2022, pp. 1–8

doi: DOI HERE
Advance Access Publication Date: Day Month Year
Paper

PAPER

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,*]

[1]College of Artificial Intelligence, Nanjing Agricultural University, 1st WeiGang, Nanjing, 210095, Jiangsu, China, [2]College of Plant Protection, Nanjing Agricultural University, 1st WeiGang, Nanjing, 210095, Jiangsu, China, [3]College of Agriculture, Nanjing Agricultural University, 1st WeiGang, Nanjing, 210095, Jiangsu, China and [4]StarHelix Inc, 88 Jiangmiao Road, Nanjing, 211899, Jiangsu, China
[*]Corresponding author. Huang:jim@njau.edu.cn, Qian:glqian@njau.edu.cn

Abstract

Motivation: 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. Numerous studies have been conducted on systematically discovering AMPs from natural metagenome. Particularly, in recent years, with the advances of artificial intelligence technology, several computational models have been developed to identify potential AMPs. However, given the computational challenges posed by the short lengths of AMPs and the increasing needs for more accurate predictions, it is important to enhance AMP prediction using more comprehensive information of peptides with state-of-the-art deep learning algorithms.
Results: 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 1900-dimension feature vector inferred by an mLSTM model 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.
Availability: UniAMP can be accessed online via https://amp.starhelix.cn, and the source code, data, and models used in this study are available on https://github.com/quietbamboo/UniAMP.
Contact: jim@njau.edu.cn

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(Petrosillo, 2020). 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(Organization et al., 2019). Specifically, *P. aeruginosa*'s extensive number of virulence factors enable remarkable adaptability, facilitating chronic infections by tailoring its response to diverse environmental stressors(Jurado-Martín et al., 2021). Furthermore, *Candida* species are among the most common causes of invasive mycotic disease, with *Candida albicans* reigning as the leading cause of invasive candidiasis(Lee et al., 2020). AMR transmission in agriculture involves not only foodborne pathogens but also commensals and environmental microbes, posing risks to human health from animal and plant-based foods(Thanner et al., 2016). 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(Årdal et al., 2020). As a result, tackling with AMR has posed a tremendous challenge.

Developing medicines and pesticides 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, targeting infectious bacterial, viruses, and fungi(Zhang et al., 2021). Therefore, AMPs serve as a promising therapeutic option, ubiquitous in the innate immune systems of various life forms(Wang et al., 2017). In contrast to typical antibiotics, most AMPs do not hinder peptidoglycan synthesis through protein binding, instead, they form complexes with precursor molecules in the membrane, creating pores, which lead to a lower likelihood for antimicrobial resistance to develop(Boparai and Sharma, 2020).

Presently, 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(Ramazi et al., 2022). Most of these AMP databases contain antimicrobial targets of the AMPs and whether the AMPs is natural or synthetic. Researchers can consult these databases and obtain AMP-related information accordingly. However, due to the time-consuming and labor-intensive nature of high-throughput experiments for evaluating each individual AMP, 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 in the hundreds, not to mention that there are a large number of duplicate AMP entries among different databases(Porto et al., 2017). This leads to obstables for speeding up the research and application of AMPs, while in the same time indicating that a huge number of potential AMPs are 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(Agrawal et al., 2021) (Support Vector Machine), AmpGram(Burdukiewicz et al., 2020) (Random Forest), and TP-MV(Yan et al., 2022) (ensemble ML method). In recent years, there are also 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(Yan et al., 2020) (CNN), sAMP-PFPDeep(Hussain, 2022) (CNN), and AMPlify(Li et al., 2022a) (LSTM). Most of these methods directly make predictions purely based on the amino acid sequence of a candidate AMP.

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(Hussain, 2022) 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(Chou, 2001) feature to predict AMPs(Yan et al., 2020). In particular, the increasing emphasis on predicting AMPs using structural properties arises from the notably accurate predictions of protein structures by AlphaFold(Jumper et al., 2021) and trRosetta(Du et al., 2021), for example, sAMPpred-GAT uses the peptide structures predicted by trRosetta to predict the AMPs based on the GAT(Yan et al., 2023).

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 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. In this study, we propose an AMP prediction framework, UniAMP, and evaluated its performance for predicting AMPs using Unified Representation (UniRep(Alley et al., 2019)) of peptides which represents a set of features inferred automatically by a deep learning model. At the core of UniAMP, we designed a novel deep neural network as a predictor, composed of fully connected modules and self-attention mechanisms.

In order for fair and comprehensive comparison, we aggregate benchmark datasets consisting of all the AMP entries from CAMPR4, DBAASPv3, dbAMP2, DRAMP3, LAMP2 and YADAMP, with 7366 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 F-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 greatly boost the research and discovery of AMPs, and we make UniAMP publically available (https://amp.starhelix.cn).

Materials and methods

Data collection and Dataset Preparation

AMP and Non-AMP

We collected AMP data from six public AMP databases: CAMPR4(Gawde et al., 2023), DBAASPv3(Pirtskhalava et al., 2021), dbAMP2(Jhong et al., 2022), DRAMP3(Shi et al., 2022), LAMP2(Ye et al., 2020) and YADAMP(Piotto et al., 2012). These six public AMP databases were merged into a larger database as our AMPs database, and only the experimentally valid non-duplicate data were retained. Only the sequences and antimicrobial activity information of peptides were retained in this database. The antimicrobial activity information is a list of Key-Value pairs of Target-Minimum Inhibitory Concentration(MIC), with all MIC units converted to μg/ml. Furthermore, the largest MIC was used as the unique value if the MIC records in different databases are inconsistent. Since the selection of MIC(Wang et al., 2021) as well as the length of a peptide sequence(Ma et al., 2022) had proved to be crucial, peptides with antimicrobial activity against *P. aeruginosa* and *C. albicans*, with a MIC less than 100 μg/ml, and with a length between 6 and 50 were screened out as two independent positive datasets. Eventually, the number of positive sequences for *P. aeruginosa* and *C. albicans* were 4821 and 2545 respectively.

On the other hand, we collected a total of 2,835,190 Non-AMP sequences as a negative dataset from Uniprot(Consortium,

2019) by setting 'length:[6 TO 50] NOT antimicrobial NOT antibiotic NOT antiviral NOT antifungal NOT fungicide NOT secreted NOT secretory NOT excreted NOT effector NOT defensin' as the search condition. After comparison with the AMPs database, ten sequences which have antimicrobial activity were removed from the negative dataset. It was worth noting that all sequences we collected contain only 20 canonical occurring amino acids.

Positive and Negative datasets

In order for fair comparison and the assessment of the models' robustness, we used the Cluster Database at High Identity with Tolerance (CD-HIT)(Fu et al., 2012) program with the parameters set as '-c 0.4' which means the similarity or sharing of the peptide sequences in different clusters does not exceed 40%(Veltri et al., 2018). Specifically, the positive sequences for *P. aeruginosa* and *C. albicans* were divided into 291 and 230 clusters respectively, and sequences in distinct clusters were regarded as dissimilar. In each positive dataset, all clusters were first randomly divided into training and test datasets at a ratio of 8:2. Because of the unbalanced number of sequences in each cluster, random division was performed multiple times until the number of sequences in two sets also maintained an approximate 8:2 ratio. So far, the positive sequences were divided into four sets (Table 1).

For the reason that the number of negative sequences was significantly more the number of positive sequences, only a small proportion of negative sequences was used. After splitting the negative dataset into 234,148 clusters using CD-HIT and randomly dividing these clusters into training and test datasets, the negative dataset was split into four sets (Table 1). A rule was that the number of negative sequences is 50 times that of positive sequences in the training set, and this ratio increases to 100 times in the test set. The reason for this is that AMPs are not commonly found in proteins in general, as they often contain a specific composition of amino acids, such as a balance of hydrophilic and hydrophobic amino acids, which facilitates their interaction with bacterial membranes(Boparai and Sharma, 2020). Besides, the length distribution was maintained to be similar to that of the positive sequences when selecting negative sequences (but short negative sequences are still missing even if they are all selected).

Benchmark datasets

We constructed two benchmark datasets (Table 1) for AMPs targeting *P. aeruginosa* and *C. albicans* respectively to test the performance compared with previous methods. Specifically, all positive sequences in test datasets were selected, and the same number of negative sequences with the same length distribution were randomly selected. Since there are specific requirements of some tools, such as sequence length less than 30(Yan et al., 2020) and more than 40(Yan et al., 2023), corresponding adjustments were made according to these requirements to obtain its true performance.

Feature Extraction

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(Chou, 2001), CT(Shen et al., 2007), and AC(Zhang et al., 2019). Thirdly, all peptide features were extracted by the Unified Representation (UniRep)(Alley et al., 2019), which computes a 1900-dimensional vector containing a semantically rich, structural, evolutionary, and biophysically grounded statistical representation of an amino acid sequence.

Table 1. All datasets in this study.

Datasets	Positives	Negatives
P. aeruginosa training dataset	3828	191,400
P. aeruginosa test dataset	993	99,800
C. albicans training dataset	2036	101,800
C. albicans test dataset	509	50,900
P. aeruginosa benchamark dataset	993	993
C. albicans benchmark dataset	509	509

Note: Both positive and negative sequences were filtered by CD-HIT, and the similarity between training samples and test samples is <40%. The length distributions of positive and negative sequences in the same dataset are similar.

Pseudo Amino Acid Composition (PseAAC)

PseAAC 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(Chen et al., 2019). The encoding of PseAAC combines the hydrophobicity, hydrophilicity, and side-chain mass of amino acids. In this study, we took the number of sequence correlation factors as 4 and the weight factor for the sequence order effect as 0.05.

Conjoint Triad (CT)

The CT(Shen et al., 2007) 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 ($7 \times 7 \times 7$), 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)(Zhang et al., 2019). 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, this matrix is normalized and recalculated according to the formula. 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.

Inferred information vector

Previously, sAMPpred-GAT(Yan et al., 2023) achieved excellent results in predicting AMPs using the structural information inferred from trRosetta(Du et al., 2021). Although the studie(Wei et al., 2021) demonstrated the structural

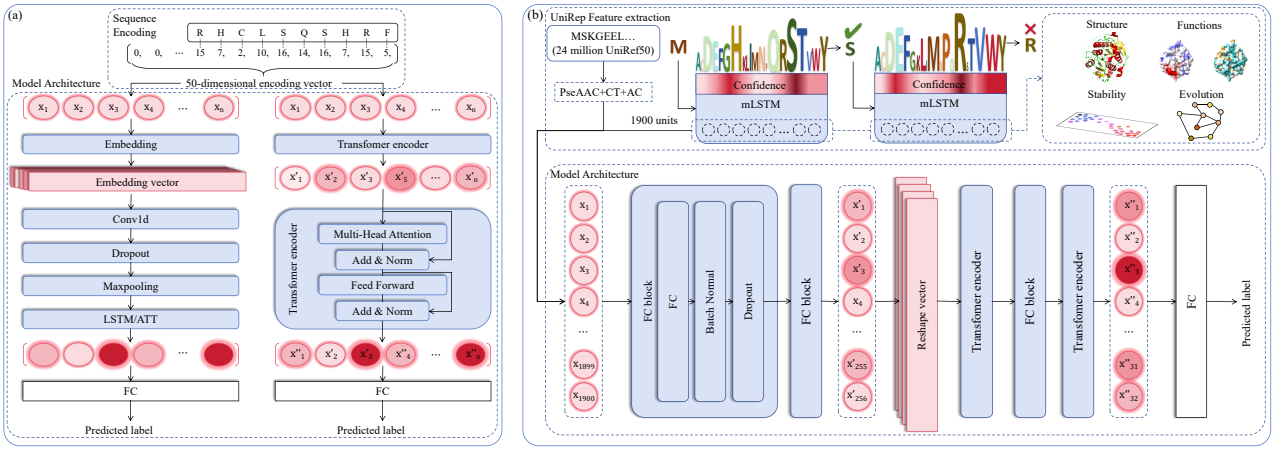

Fig. 1. The framework of the predictor in this study. (a) The workflow of the sequence models. We used the word2vec to encode the peptide sequence into a 50-dimensional vector. Modifications to the LSTM and ATT models were limited to a single network layer, with the remaining network architecture kept consistent. BERT was constructed based on the transformer encoder. Besides evaluating three sequence models, we took the intersection of the three predictions as an additional comparison with the conclusions from the study(Ma et al., 2022). In the figure, the deeper shades of red signify richer information. (b) The workflow of the UniAMP. We employed UniRep for feature extraction. The mLSTM model for amino acid sequence prediction incorporates 1900 hidden units encompassing information on protein structure, function, stability, evolution, and so on. In addition to the 1900-dimensional vector, we incorporated PseAAC, CT, and AC, concatenating them into a 402-dimensional vector for comparison. The prediction model maps inputs to a fixed length using two enhanced fully connected layers. Subsequently, two transformer encoders facilitates inter-feature information transfer, generating context-rich feature vector. And prediction labels are eventually output.

similarity of AMPs with the same functions, the structural information is obviously not comprehensive enough for predicting AMPs. Therefore, we used the comprehensive information of peptides inferred by the UniRep(Alley et al., 2019) in this study.

The UniRep, trained on 24 million UniRef50(Suzek et al., 2015) amino acid sequences using a 1,900-hidden unit Multiplicative long-/short-term-memory (mLSTM) RNNs model capable of fully learning the rich information of natural language to generate protein sequences(Radford et al., 2017), exhibits several capabilities (Figure 1). 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, The 1900-dimensional UniRep vector encapsulates not only composition, physicochemical and structural properties, but also a wealth of information, encoding structural, evolutionary, and functional insights.

Based on a trained UniRep model, we converted the peptide into a 1900-dimensional vector as the input to the model (Figure 1), and believed that this inferred vector contained comprehensive information.

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(Ma et al., 2022). 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(Veltri et al., 2018) combined with the NLP algorithm had proved effective(Ma et al., 2022). In particular, this method performs well on datasets containing a substantial number of negative sequences, with precision several times than previous approaches. As part of this analysis, three NNM based on NLP algorithms were used for sequence-based AMPs prediction.

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 ATT layer, while the third model was BERT model based on transformer encoders(Vaswani et al., 2017; Devlin et al., 2018) (Figure 1). Like training the NLP model, we treat the amino acid sequence as a 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 (Figure 1).

Feature vector Model

When designing the model, given the relatively comprehensiveness of the extracted feature information, our focus was on using fully connected layers to achieve a more effective classification of feature information. Consequently, drawing inspiration from the study(Li et al., 2022b), we devised a network architecture incorporating two transformer encoders. Our objective was to facilitate the transmission of information between features, particularly over longer distances, using the self-attention

mechanism of the transformer encoder, which is difficult to achieve with fully connected layers. Additionally, the incorporation of batch normalization and dropout layers within the fully connected module had been undertaken to enhance the stability of the model during feature mapping, and to mitigate issues like overfitting, gradient vanishing, and exploding.

The model architecture is shown in Figure 1. The feature vectors are mapped to a 256-dimensional vector through two fully connected modules. Subsequently, the reshaped vectors are input into two transformer encoders, and the predicted labels are then generated using fully connected layers.

Performance measure

Evaluation metrics

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

$$\begin{cases} \text{Accuracy} = \dfrac{TP+TN}{TP+FP+TN+FN} \\ \text{Precision} = \dfrac{TP}{TP+FP} \\ \text{Recall} = \dfrac{TP}{TP+FN} \\ \text{F}_1\text{-score} = \dfrac{(1+\beta^2)\times \text{Precision}\times \text{Recall}}{\beta^2\times \text{Precision}+\text{Recall}} \\ \text{MCC} = \dfrac{TP\times TN - FP\times FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}} \\ \text{AUC : Area Under the ROC Curve} \end{cases} \tag{1}$$

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 classification performance, which is especially valuable in scenarios with imbalanced class distribution(Chicco and Jurman, 2020).

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=0.0001). 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. Besides, in training the models proposed in this paper, each model was trained five times using the previously mentioned training methods rigorously, and the model with the median MCC performance was selected for comparison.

Results

Performance comparison on benchmark datasets

To accurately assess the performance of UniAMP, we concurrently evaluated multiple predictors on the benchmark datasets (Table 2), including CAMPR4(Gawde et al., 2023), AmPEP(Bhadra et al., 2018; Yan et al., 2020), amPEPpy(Lawrence et al., 2021), AMPfun(Chung et al., 2020), AmpGram(Burdukiewicz et al., 2020), AMPScannerV2(Veltri et al., 2018), and sAMPpred-GAT(Yan et al., 2023). It is worth noting that *P. aeruginosa* falls under the category of Gram-negative bacteria, and *C. albicans* is an infectious fungus. Although these predictors not only forecast AMPs targeting *P. aeruginosa* and *C. albicans*, all positive samples in benchmark datasets all belong to the AMP category, whereas negative samples exhibit no antimicrobial activity and belong to the non-AMP category. This aligns with the requirements of the predictors. Furthermore, a portion of the benchmark datasets is sourced from CAMPR4, with some overlap with data used by other methods (it is unclear whether this subset was used for training) As a result, the performance of existing predictors, especially CAMPR4, may be overestimated.

The performance of each predictor is presented in Table 2a. UniAMP exhibits the highest accuracy, precision, F_1-score, and MCC on both benchmark datasets, and the highest recall among the five models trained in this study. When the feature extraction of UniAMP was replaced with manual extraction, its comprehensive performance notebly declined, positioning it only in the mid-range among existing tools. Furthermore, sAMPpred-GAT uses the structural information inferred by deep learning models for prediction and achieves good performance, attaining the highest recall and AUC among all predictors. The evaluation results show that the information inferred by deep learning models can improve prediction performance, and UniAMP is a comprehensive predictor.

Performance comparison on test datasets

The performance is presented in Table 2b. The approach of intersecting predictive models significantly improved precision, exhibiting markedly higher precision across both test datasets compared to alternative methods. However, UniAMP achieved the highest accuracy (minimal sum of FP and FN), recall, F_1-score, and MCC. Although precision is an important metric in the AMPs discovery process, during model training, we filtered models based on MMC as primary evaluation metric. It is worth noting that LSTM achieved optimal performance on the validation dataset (only leading by 0.0003), whereas on the test dataset, UniAMP exhibited superior performance (leading by 0.0147). Moreover, the MCC of UniAMP surpasses that of the balanced benchmark datasets when evaluated on imbalanced testing datasets. The results indicate that UniAMP exhibits robustness and maintains strong performance even in imbalanced real-world scenarios.

Feature extraction

In this section, We evaluate various combinations of manual feature extraction methods, and thier performance on *P. aeruginosa* datasets is reported in Table 3. 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

Table 2. Performance of UniAMP and some existing AMPs predictors on benchmark datasets and test datasets.

(a) *P. aeruginosa* benchmark dataset

Method	TP	FP	TN	FN	Acc	Pre	Rec	F_1-score	MCC
CAMPR4-RF(Gawde et al., 2023)	970	167	826	23	0.9043	0.8531	0.9768	0.9108	0.8173
CAMPR4-SVM(Gawde et al., 2023)	935	142	851	58	0.8993	0.8682	0.9416	0.9034	0.8015
RF-AmPEP30[a](Yan et al., 2020)	851	137	667	43	0.894	0.8613	0.9519	0.9043	0.7911
AMPScannerV2(Veltri et al., 2018)	917	172	821	56	0.8852	0.8449	0.9436	0.8915	0.7757
sAMPpred-GAT[a](Yan et al., 2023)	48	19	94	0	0.882	0.7164	**1.0**	0.8348	0.772
amPEPpy(Lawrence et al., 2021)	927	172	821	66	0.8802	0.8435	0.9335	0.8862	0.7647
AmpGram[a](Burdukiewicz et al., 2020)	808	202	660	54	0.8515	0.800	0.9374	0.8633	0.7136
CAMPR4-ANN(Gawde et al., 2023)	837	148	845	156	0.8469	0.8497	0.8463	0.8484	0.6939
AMPfun(Chung et al., 2020)	940	366	627	53	0.789	0.7198	0.9466	0.8178	0.6091
AmPEP(Bhadra et al., 2018)	544	418	575	449	0.5634	0.5655	0.5478	0.5565	0.127
LSTM	826	3	990	167	0.9144	0.9964	0.8318	0.9067	0.8403
ATT	826	5	988	167	0.9133	0.994	0.8318	0.9057	0.838
BERT	848	4	989	145	0.925	0.9953	0.854	0.9193	0.8586
UniAMP (with PseAAC, CT, and AC)	771	6	987	222	0.8852	0.9923	0.7764	0.8712	0.7893
UniAMP	848	1	992	145	**0.9265**	**0.9988**	0.854	**0.9207**	**0.8621**

(b) *P. aeruginosa* test dataset[b]

Method	TP	FP	TN	FN	Pre	Rec	F_1-score	MCC
LSTM	826	41	99259	167	0.9527	0.8318	0.8882	0.8892
ATT	826	77	99223	167	0.9147	0.8318	0.8713	0.8711
BERT	848	94	99206	145	0.9002	**0.854**	0.8765	0.8756
LSTM&ATT&BERT	760	10	99290	233	**0.987**	0.7654	0.8622	0.8681
UniAMP (with PseAAC, CT and AC)	771	124	99176	222	0.8615	0.7764	0.8168	0.8161
UniAMP	848	61	99239	145	0.9329	**0.854**	**0.8917**	**0.8915**

Note: Existing AMPs Predictors used trained models published by the creators. For a fair comparison, 5 times of each model in this article were trained, and the model with the median MCC performance was selected for comparison.

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

[b]Test datasets does not employ accuracy as an evaluation metric due to the abundance of negative samples.

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(Liu and Gillies, 2016). 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. On the contrary, inferred information demonstrated consistently high performance on both the validation and test datasets, attesting to its robustness.

Inferred information

Among the five models used in this study, LSTM exhibits the closest performance to UniAMP, and the UniRep feature vectors are derived from the hidden units of mLSTM. Subsequently, we compared the output of 100 hidden units from the LSTM model with the 1900-dimensional vector of UniRep. To our surprise, we observed a significant correlation (Pearson's r=0.47 $p<6.08\times10^{-7}$) between the LSTM vector and a 100-dimensional segment starting from the 146th dimension of the UniRep vector. As the UniRep vector has been previously validated to be rich in protein-related information(Alley et al., 2019), we could believe that LSTM, as a sequence model, has learned knowledge beyond the sequence through its feature extraction module.

Table 3. Performance of combined manual feature extraction.

Method	MCC (validation)	MCC (test)
PseAAC	0.8879	**0.8315**
CT	0.8357	0.7217
AC	0.772	0.6968
PseAAC+CT	0.8864	0.7877
PseAAC+AC	0.8856	0.8197
CT+AC	0.8802	0.7997
PseAAC+CT+AC	**0.8936**	0.8161

Note: Because the patience value of 30, we report the performance of each model for the first 30 epochs before the training cessation.

Discussion

Previously, most predictors were designed for all AMPs, providing them with a larger dataset for training. However, when targeting specific pathogens, such as E.coli(Wang et al., 2021), the available positive data is only one-tenth of the entire dataset, making model training more challenging. The

feature vectors utilized by UniAMP were derived from a protein-generating mLSTM model, indicating that its feature extraction is independent on the limited availability of positive data. Given the abundance of protein data, this approach allows us to extract information related to protein structure and function. The surprising correlation between the intermediate output of LSTM and the input features of UniAMP confirms that the inferred information utilized by UniAMP incorporates the information extracted by the sequence model. These two vectors, serving as the outputs of the 'hidden layer' before the fully connected layer, are commonly regarded as highly abstracted fusion features of the input data(Zeiler and Fergus, 2014). Such fusion features often exhibit enhanced adaptability and generalization capabilities. Experimental results further indicate that a subset of this information indeed plays a crucial role in determining peptide functionality. This has inspired us to consider that inferred information obtained from models with commonalities, such as protein-related generation, structure prediction, functional prediction, may make up for data scarcity.

The sAMPpred-GAT achieves the highest AUC and recall in the evaluation results. However, sAMPpred-GAT confines the inferred information to structural aspects, which may be not comprehensive enough for predicting AMP. This may be the reason its comprehensive performance is not as good as UniAMP. This insight emphasizes the significance of comprehensiveness in feature extraction, whether derived through manual or deep learning methods. Additionally, employing the method of taking the intersection of different models indeed improves precision. Although it incurs a significant loss in recall (with a precision increase of less than 0.1 and a recall decrease exceeding 0.13), resulting in an overall performance decline, it maybe highly effective in scenarios where precision is of paramount importance.

Future research could explore not only similar approaches for feature extraction to attain higher model performance but also consider establishing a mapping between manual features and deep learning features based on prior knowledge. This is akin to reasoning backward from a favorable outcome to understand the underlying reasons, thereby aiding in refining the complete functional mechanisms of AMPs. In summary, we hope that UniAMP can not only serve as an excellent AMPs predictor, but also provide a novel AMPs research perspective: utilizing information inferred by deep learning models.

Conclusion

In this study, we proposed a framework for predicting AMPs called UniAMP. This approach utilized the output of 1900-hidden units from an mLSTM model designed for protein sequence generation as inferred information for peptides. This information was input into a prediction network composed of fully connected modules and a transformer encoder to predict antibacterial activity. Evaluation results demonstrate that the model achieves the best comprehensive performance on both a balanced benchmark dataset and an imbalanced test dataset. We analyze the relationship between peptide sequences, manually extracted features, and inferred information, ultimately concluding that the inferred information from deep learning models is more comprehensive and non-redundant. This characteristic contributes to UniAMP's excellent performance and robustness, and this approach exhibits potential applications in future research.

Data and Code availability

The source code, data, and models used in this study are available on https://github.com/quietbamboo/UniAMP. The complete dataset of non-AMP data can be downloaded from Uniprot based on the conditions specified in the article. And providing all relevant links for AMPs:CAMPR4, DBAASPv3, dbAMP2, DRAMP3, LAMP2, and YADAMP

References

P. Agrawal, D. Bhagat, M. Mahalwal, N. Sharma, and G. P. Raghava. Anticp 2.0: an updated model for predicting anticancer peptides. *Briefings in bioinformatics*, 22(3): bbaa153, 2021.

E. C. Alley, G. Khimulya, S. Biswas, M. AlQuraishi, and G. M. Church. Unified rational protein engineering with sequence-based deep representation learning. *Nature methods*, 16(12): 1315–1322, 2019.

C. Årdal, M. Balasegaram, R. Laxminarayan, D. McAdams, K. Outterson, J. H. Rex, and N. Sumpradit. Antibiotic development—economic, regulatory and societal challenges. *Nature Reviews Microbiology*, 18(5):267–274, 2020.

P. Bhadra, J. Yan, J. Li, S. Fong, and S. W. Siu. Ampep: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. *Scientific reports*, 8(1):1697, 2018.

J. K. Boparai and P. K. Sharma. Mini review on antimicrobial peptides, sources, mechanism and recent applications. *Protein and peptide letters*, 27(1):4–16, 2020.

M. Burdukiewicz, K. Sidorczuk, D. Rafacz, F. Pietluch, J. Chilimoniuk, S. Rödiger, and P. Gagat. Proteomic screening for prediction and design of antimicrobial peptides with ampgram. *International journal of molecular sciences*, 21(12):4310, 2020.

C. Chen, Q. Zhang, Q. Ma, and B. Yu. Lightgbm-ppi: Predicting protein-protein interactions through lightgbm with multi-information fusion. *Chemometrics and Intelligent Laboratory Systems*, 191:54–64, 2019.

D. Chicco and G. Jurman. The advantages of the matthews correlation coefficient (mcc) over f1 score and accuracy in binary classification evaluation. *BMC genomics*, 21(1):1–13, 2020.

K.-C. Chou. Prediction of protein cellular attributes using pseudo-amino acid composition. *Proteins: Structure, Function, and Bioinformatics*, 43(3):246–255, 2001.

C.-R. Chung, T.-R. Kuo, L.-C. Wu, T.-Y. Lee, and J.-T. Horng. Characterization and identification of antimicrobial peptides with different functional activities. *Briefings in bioinformatics*, 21(3):1098–1114, 2020.

U. Consortium. Uniprot: a worldwide hub of protein knowledge. *Nucleic acids research*, 47(D1):D506–D515, 2019.

J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Z. Du, H. Su, W. Wang, L. Ye, H. Wei, Z. Peng, I. Anishchenko, D. Baker, and J. Yang. The trrosetta server for fast and accurate protein structure prediction. *Nature protocols*, 16 (12):5634–5651, 2021.

L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li. Cd-hit: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23):3150–3152, 2012.

U. Gawde, S. Chakraborty, F. H. Waghu, R. S. Barai, A. Khanderkar, R. Indraguru, T. Shirsat, and S. Idicula-Thomas. Campr4: a database of natural and synthetic antimicrobial peptides. *Nucleic Acids Research*, 51(D1): D377–D383, 2023.

W. Hussain. samp-pfpdeep: Improving accuracy of short antimicrobial peptides prediction using three different sequence encodings and deep neural networks. *Briefings in Bioinformatics*, 23(1):bbab487, 2022.

J.-H. Jhong, L. Yao, Y. Pang, Z. Li, C.-R. Chung, R. Wang, S. Li, W. Li, M. Luo, R. Ma, et al. dbamp 2.0: updated resource for antimicrobial peptides with an enhanced scanning method for genomic and proteomic data. *Nucleic Acids Research*, 50(D1):D460–D470, 2022.

J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

I. Jurado-Martín, M. Sainz-Mejías, and S. McClean. Pseudomonas aeruginosa: An audacious pathogen with an adaptable arsenal of virulence factors. *International journal of molecular sciences*, 22(6):3128, 2021.

T. J. Lawrence, D. L. Carper, M. K. Spangler, A. A. Carrell, T. A. Rush, S. J. Minter, D. J. Weston, and J. L. Labbé. ampeppy 1.0: a portable and accurate antimicrobial peptide prediction tool. *Bioinformatics*, 37(14):2058–2060, 2021.

Y. Lee, E. Puumala, N. Robbins, and L. E. Cowen. Antifungal drug resistance: molecular mechanisms in candida albicans and beyond. *Chemical reviews*, 121(6):3390–3411, 2020.

C. Li, D. Sutherland, S. A. Hammond, C. Yang, F. Taho, L. Bergman, S. Houston, R. L. Warren, T. Wong, L. M. Hoang, et al. Amplify: attentive deep learning model for discovery of novel antimicrobial peptides effective against who priority pathogens. *BMC genomics*, 23(1):77, 2022a.

X. Li, P. Han, G. Wang, W. Chen, S. Wang, and T. Song. Sdnn-ppi: self-attention with deep neural network effect on protein-protein interaction prediction. *BMC genomics*, 23 (1):474, 2022b.

R. Liu and D. F. Gillies. Overfitting in linear feature extraction for classification of high-dimensional image data. *Pattern Recognition*, 53:73–86, 2016.

Y. Ma, Z. Guo, B. Xia, Y. Zhang, X. Liu, Y. Yu, N. Tang, X. Tong, M. Wang, X. Ye, et al. Identification of antimicrobial peptides from the human gut microbiome using deep learning. *Nature Biotechnology*, 40(6):921–931, 2022.

W. H. Organization et al. 2019 antibacterial agents in clinical development: an analysis of the antibacterial clinical development pipeline. 2019.

N. Petrosillo. Infections: the emergency of the new millennium. *Nuclear Medicine in Infectious Diseases*, pages 1–8, 2020.

S. P. Piotto, L. Sessa, S. Concilio, and P. Iannelli. Yadamp: yet another database of antimicrobial peptides. *International journal of antimicrobial agents*, 39(4):346–351, 2012.

M. Pirtskhalava, A. A. Amstrong, M. Grigolava, M. Chubinidze, E. Alimbarashvili, B. Vishnepolsky, A. Gabrielian, A. Rosenthal, D. E. Hurt, and M. Tartakovsky. Dbaasp v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. *Nucleic acids research*, 49(D1):D288–D297, 2021.

W. Porto, A. Pires, and O. Franco. Computational tools for exploring sequence databases as a resource for antimicrobial peptides. *Biotechnology advances*, 35(3):337–349, 2017.

A. Radford, R. Jozefowicz, and I. Sutskever. Learning to generate reviews and discovering sentiment. *arXiv preprint arXiv:1704.01444*, 2017.

S. Ramazi, N. Mohammadi, A. Allahverdi, E. Khalili, and P. Abdolmaleki. A review on antimicrobial peptides databases and the computational tools. *Database*, 2022: baac011, 2022.

J. Shen, J. Zhang, X. Luo, W. Zhu, K. Yu, K. Chen, Y. Li, and H. Jiang. Predicting protein–protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences*, 104(11):4337–4341, 2007.

G. Shi, X. Kang, F. Dong, Y. Liu, N. Zhu, Y. Hu, H. Xu, X. Lao, and H. Zheng. Dramp 3.0: an enhanced comprehensive data repository of antimicrobial peptides. *Nucleic Acids Research*, 50(D1):D488–D496, 2022.

B. E. Suzek, Y. Wang, H. Huang, P. B. McGarvey, C. H. Wu, and U. Consortium. Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932, 2015.

S. Thanner, D. Drissner, and F. Walsh. Antimicrobial resistance in agriculture. *mBio*, 7(2):e02227–15, 2016.

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

D. Veltri, U. Kamath, and A. Shehu. Deep learning improves antimicrobial peptide recognition. *Bioinformatics*, 34(16): 2740–2747, 2018.

C. Wang, S. Garlick, and M. Zloh. Deep learning for novel antimicrobial peptide design. *Biomolecules*, 11(3):471, 2021.

P. Wang, R. Ge, L. Liu, X. Xiao, Y. Li, and Y. Cai. Multi-label learning for predicting the activities of antimicrobial peptides. *Scientific reports*, 7(1):2202, 2017.

L. Wei, W. He, A. Malik, R. Su, L. Cui, and B. Manavalan. Computational prediction and interpretation of cell-specific replication origin sites from multiple eukaryotes by exploiting stacking framework. *Briefings in Bioinformatics*, 22(4): bbaa275, 2021.

J. Yan, P. Bhadra, A. Li, P. Sethiya, L. Qin, H. K. Tai, K. H. Wong, and S. W. Siu. Deep-ampep30: improve short antimicrobial peptides prediction with deep learning. *Molecular Therapy-Nucleic Acids*, 20:882–894, 2020.

K. Yan, H. Lv, J. Wen, Y. Guo, and B. Liu. Tp-mv: therapeutic peptides prediction by multi-view learning. *Current Bioinformatics*, 17(2):174–183, 2022.

K. Yan, H. Lv, Y. Guo, W. Peng, and B. Liu. sampred-gat: prediction of antimicrobial peptide by graph attention network and predicted peptide structure. *Bioinformatics*, 39(1):btac715, 2023.

G. Ye, H. Wu, J. Huang, W. Wang, K. Ge, G. Li, J. Zhong, and Q. Huang. Lamp2: a major update of the database linking antimicrobial peptides. *Database*, 2020:baaa061, 2020.

M. D. Zeiler and R. Fergus. Visualizing and understanding convolutional networks. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I 13*, pages 818–833. Springer, 2014.

L. Zhang, G. Yu, D. Xia, and J. Wang. Protein–protein interactions prediction based on ensemble deep neural networks. *Neurocomputing*, 324:10–19, 2019.

Q.-Y. Zhang, Z.-B. Yan, Y.-M. Meng, X.-Y. Hong, G. Shao, J.-J. Ma, X.-R. Cheng, J. Liu, J. Kang, and C.-Y. Fu. Antimicrobial peptides: mechanism of action, activity and clinical potential. *Military Medical Research*, 8:1–25, 2021.