Journal of Advanced Research xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of Advanced Research

journal homepage: www.elsevier.com/locate/jare

Original Manuscript

AMPCliff: Quantitative definition and benchmarking of activity cliffs in antimicrobial peptides

Kewei Li [a], Yuqian Wu [b], Yinheng Li [d], Yutong Guo [c], Yanwen Kong [a], Yan Wang [g], Yiyang Liang [h], Yusi Fan [a], Lan Huang [a], Ruochi Zhang [e,*], Fengfeng Zhou [a,f,*]

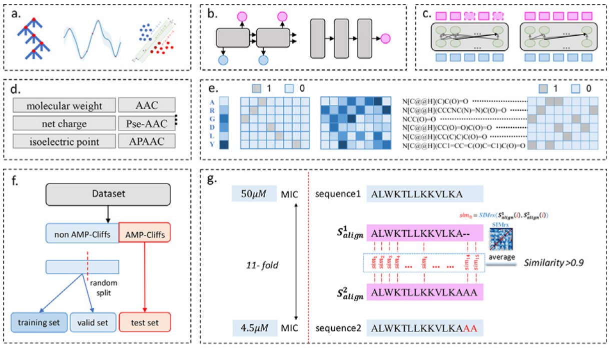
[a] College of Computer Science and Technology, and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin 130012, China
[b] School of Software, Jilin University, Changchun 130012 Jilin, China
[c] School of Life Sciences, Jilin University, Changchun 130012 Jilin, China
[d] Department of Computer Science, Columbia University, 116th and Broadway, New York City, NY 10027, United States
[e] School of Artificial Intelligence, and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun 130012 Jilin, China
[f] School of Biology and Engineering, Guizhou Medical University, Guiyang 550025 Guizhou, China
[g] School of Computer Engineering, Changchun University of Engineering, Changchun 130103 Jilin, China
[h] Changchun Wenli High School, Changchun 130062 Jilin, China

HIGHLIGHTS

- We present the first systematic screening of the AC Phenomenon in AMPs, called AMPCliff.
- We propose AMPCliff-specific data partition (AC split) and suitable evaluation metrics.
- We conduct a comprehensive benchmark of ML and DL models for AMPCliff prediction task.

GRAPHICAL ABSTRACT

An overview of the quantitative definition and benchmarking of AMPCliff. a-c are the model architectures used in this paper, including **a** machine learning methods: RF, XGBoost, GB, GP and SVM. **b** deep learning methods LSTM and CNN. **c** pre-trained language models like transformer encoder-based models BERT, ESM2 and transformer decoder-based model GPT2, ProGen2. **d-e** list the features used in this paper, including **d** fixed type representations, **e** dictionary index, one-hot encoding, word embedding and fingerprint. **f** the procedure of AC split. **g** an illustrative definition of AMPCliff.



ARTICLE INFO

Article history:
Received 31 January 2025
Revised 9 April 2025
Accepted 29 April 2025
Available online xxxx

ABSTRACT

Introduction: Activity cliff (AC) is a phenomenon that a pair of similar molecules differ by a small structural alternation but exhibit a large difference in their biochemical activities. This phenomenon affects various tasks ranging from virtual screening to lead optimization in drug development. The AC of small molecules has been extensively investigated but limited knowledge is accumulated about the AC phenomenon in pharmaceutical peptides with canonical amino acids.

* Corresponding authors at: College of Computer Science and Technology, and Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education, Jilin University, Changchun, Jilin 130012, China (Fengfeng Zhou).

E-mail addresses: zrc720@gmail.com (R. Zhang), ffzhou@jlu.edu.cn, FengfengZhou@gmail.com (F. Zhou).

https://doi.org/10.1016/j.jare.2025.04.046

Please cite this article as: K. Li, Y. Wu, Y. Li et al., AMPCliff: Quantitative definition and benchmarking of activity cliffs in antimicrobial peptides, Journal of Advanced Research, https://doi.org/10.1016/j.jare.2025.04.046

Keywords:
AMPCliff
Activity cliff (AC)
Minimum inhibitory concentration (MIC)
Benchmark
Antimicrobial peptide (AMP)

Objectives: This study introduces a quantitative definition and benchmarking framework AMPCliff for the AC phenomenon in antimicrobial peptides (AMPs) composed by canonical amino acids.
Methods: This study establishes a benchmark dataset of paired AMPs in *Staphylococcus aureus* from the publicly available AMP dataset GRAMPA, and conducts a rigorous procedure to evaluate various AMP AC prediction models, including nine machine learning, four deep learning algorithms, four masked language models, and four generative language models.
Results: A comprehensive analysis of the existing AMP dataset reveals a significant prevalence of AC within AMPs. AMPCliff quantifies the activities of AMPs by the metric minimum inhibitory concentration (MIC), and defines 0.9 as the minimum threshold for the normalized BLOSUM62 similarity score between a pair of aligned peptides with at least two-fold MIC changes. Our analysis reveals that these models are capable of detecting AMP AC events and the pre-trained protein language model ESM2 demonstrates superior performance across the evaluations. The predictive performance of AMP activity cliffs remains to be further improved, considering that ESM2 with 33 layers only achieves the Spearman correlation coefficient 0.4669 for the regression task of the –log(MIC) values on the benchmark dataset.
Conclusion: Our findings highlight limitations in current deep learning–based representation models. To more accurately capture the properties of antimicrobial peptides (AMPs), it is essential to integrate atomic-level dynamic information that reflects their underlying mechanisms of action.

Introduction

Activity cliff (AC) refers to a pair of similar compounds that differ by only a small structural change but exhibit a large difference in their biochemical activities [1–11]. AC was first observed in the investigations of Quantitative Structure-Activity Relationship (QSAR) modeling in drug design [12], and has remained a persistent challenge in QSAR modeling for over 30 years [1]. This phenomenon affects tasks ranging from virtual screening to lead optimization in drug development [9,13]. Precise AC modeling is highly beneficial for pharmaceutical researchers.

In earlier studies, researchers designed hand-crafted features and linear functions to construct SAR models, such as HomoSAR [14]. However, these approaches could only capture simple relationships and were insufficient for modeling complex, nonlinear interactions among features. Advances in deep learning have significantly driven the development of drug design and the applications of QSAR modeling in scientific research and industry [15–17]. Recent efforts in QSAR modeling of antimicrobial peptides (AMP) include Recurrent Neural Network (RNN) models [18] and generative variational autoencoder (VAE) [19]. Despite these advancements, the best models have only achieved a Pearson correlation coefficient (PCC) of 0.77. This observation suggests that ACs could be a critical factor in building precise QSAR models for AMPs.

The early studies on the activity of antimicrobial peptides have already observed phenomena similar to activity cliffs. The experiments in this paper[20] found that after deleting two bases from the N-terminal of VFRLKKWMQKVIDRFGG, the MIC against *Staphylococcus aureus* (*S. aureus*) increased from 32 μM to above 128 μM [20]. Another study [21] observed that the C-terminal regions of porcine thrombin showed antimicrobial activities against both Gram-positive and Gram-negative bacteria. It seems to be aligned with the general trend that the antimicrobial activity of the peptides generally correlated positively with the length of the peptide found in this work [22]. At the same time, they observed some special cases. Changing WRWWWR peptide against *S. aureus* to WRWWWRW had an IC50 value of approximately 32 μM increased to above 100 μM [22]. RWWWW still had an IC50 around 32 μM, while RWWWWW completely lost its activity [22]. On the other hand, some indels can also result in AC. The AMP in [23] showed a dramatic loss of activity in both L → V and L → B variants. And the work [24] observed the same phenomenon. The activity of R2AW → R2AW(1–22) decreased more than 10 times. Although these findings explain this unusual phenomenon, it is difficult to accurately capture. It exists not only in AMPs, but also in the small molecule drug design field, which has led researchers to study this phenomenon more systematically.

AC studies in small molecules can be separated into three types (Fig. 1): (i) discovery, (ii) data-driven analysis, and (iii) predictive modeling. Since 2006, the concept of AC received multiple precise definitions [3–5,9,25], such as MMP-Cliff [3] and 3D activity cliffs [5]. Various metrics [2,7,26] like SARI [26] and SALI [2] have also been proposed to characterize ACs. The term MMP-Cliff was defined to focus on structural differences within molecule pairs in 2012 [3], and remains the most prevalent form in the AC characterizations. The same team also explored 3D activity cliffs [5], which highlighted the challenges of applying 2D measures to 3D structures and emphasized the importance of molecular representation in AC characterization. Despite such advances, the exploration of 3D cliffs has stalled since 2015 [8], possibly due to limited data. The popularly-used term MMP-Cliff requires two molecules in the MMP (Matched Molecular Pair) to differ at most eight non-hydrogen atoms in the exchanged fragments of the size at most 13 non-hydrogen atoms, and to have the potency difference at least 2 orders of magnitudes.

Most AC studies before 2010 focused on the data-driven analysis of what caused the AC phenomenon (Fig. 1). Afterwards, numerous predictive models based on the definition of MMP-Cliff have been developed [1,9], e.g., random forests [27] and support vector machines [28]. Deep learning (DL) algorithms have also been explored for their applications in this field and delivered limited successes [1,6] due to their inability to capture the high-frequency signals in the AC pairs [29–33]. Machine learning (ML) models with stereochemistry-aware features [6] like Extended Connectivity Fingerprints (ECFP) have proven more effective in characterizing the MMP-Cliff molecular pairs than end-to-end deep learning algorithms [6,11]. Recent research has taken a fresh look at QSAR models from an algorithmic perspective, and pinpointed fundamental issues with deep learning in predicting small molecule properties, beyond just data scarcity [31].

This study quantitatively defines the AC Phenomenon in AMP as a pair of AMPs with high sequence similarity but radically different antimicrobial activities (called an AMPCliff). We systematically evaluate the recent peptide representation algorithms on the AMPCliff prediction task. To the best of our knowledge, there is no direct quantitative definition of ACs in AMPs with canonical amino acids. We reviewed the history of AC investigations in small molecules since its concept has been extensively studied. We identify that the popularly employed definition of MMP-Cliff in small molecules fails to consider the evolutionary conservations among pep-

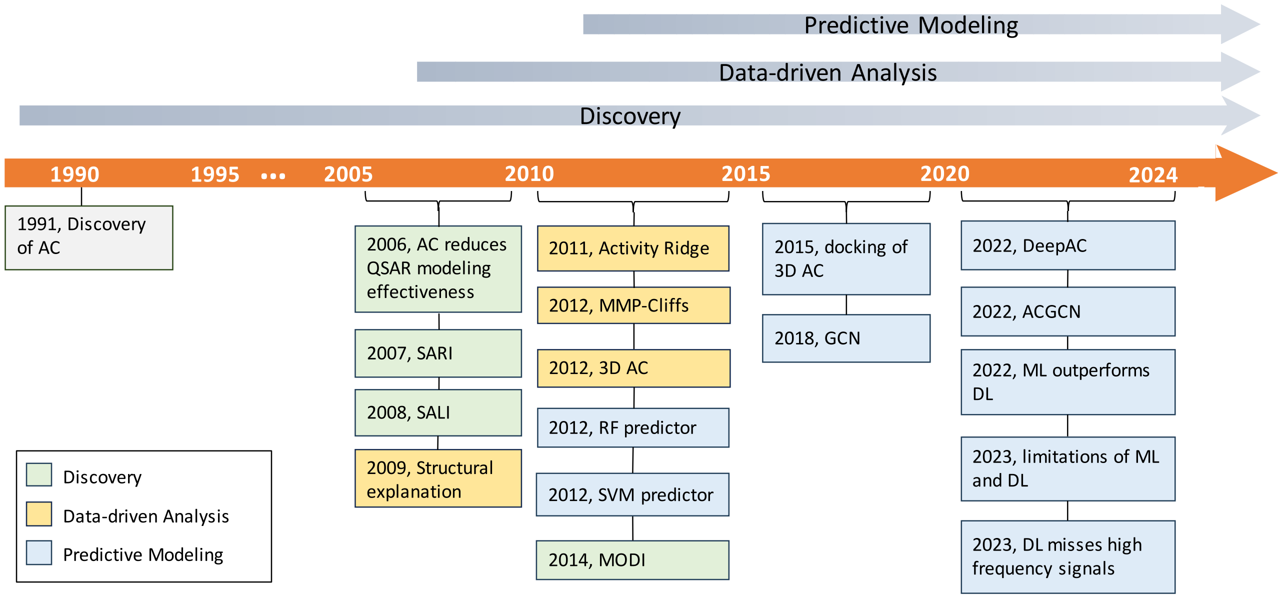


Fig. 1. A brief history of the investigations of the AC phenomenon for the molecular prediction task.

tides. It is widely accepted that two proteins or peptides with evolutionarily conserved substitutions tend to exert similar biochemical activities [34]. This oversight is understandable that evolutionary conservation is not as important in small molecules as in peptides.

The prevalence of AMPs with canonical amino acids in the public datasets constitutes the basis of quantitatively defining and benchmarking the AMPCliff pairs in this study. The main contributions of this study are summarized as follows:

- We present the first systematic screening of the AC Phenomenon in AMPs (AMPCliff), based on the quantitative activity measurement of their Minimum Inhibitory Concentrations (MICs) against the same bacterium.
- The quantitative definition of AMPCliff is proposed based on the AMPs of *S. aureus* in the public AMP dataset GRAMPA, and evaluated by a novel data partition method, AC split, for model training.
- We conduct a comprehensive benchmarking experiment of various machine learning, deep learning, and pre-trained language models on the AMPCliff prediction task.

Preliminaries

The detailed descriptions of the peptide representations and the model architectures are provided in Supplementary Material S1.

Benchmark dataset

Description of the benchmark dataset. Building a QSAR model for AMP design has been significantly hindered due to the limited availability of quantitative measurements of AMP activity, such as MIC values, in public datasets like APD3 [35], DRAMP [36], DBAASP [37], and YADAMP [38]. Witten et al. finally curated the MIC values of AMPs from these public datasets in the Spring of 2018 [39]. To the best of our knowledge, GRAMPA [39] is the first and only public AMP dataset with MIC values, and it has already facilitated the studies of peptide design. Huang et al. trained an RNN regressive model based on the GRAMPA dataset to filter highly active peptides [18]. Pandi et al. also filtered sequences against *Escherichia coli* (*E. coli*) and *Bacillus subtilis (B. subtilis)* from GRAMPA to build RNN-based and CNN-based regressive models [19]. GRAMPA contains 6760 unique sequences, and 51,345 total MIC measurements (Fig. 2**a Origin**). Some AMP/bacterium pairs occur multiple times due to overlap between databases and/or antimicrobial activity tests against multiple bacterial sub-strains [39].

We follow the data processing strategy in [39] to remove any peptides containing Cysteine for the exclusion of disulfide bonds. 35,862 entries are obtained for further analysis (Fig. 2**a w/o_Cys**). We take the geometric mean when multiple measurements for an AMP/bacterium pair are present in the database, similar as in [39], and finally get 25,628 entries (Fig. 2**a w/o_Cys_Unique**). The preprocessed dataset consists of 3759 AMPs with their associated MIC values against *E. coli*, and 3373 peptides against *S. aureus*. These two bacteria carry the rich annotations in the database. Since *S. aureus* is one of the most prevalent pathogens threatening human life for their rapid increase in antimicrobial resistance [40] (Fig. 2**b**), this study uses *S. aureus* as an example to discuss the definition and prediction of AMPCliffs. More than 80 % of the entries in the dataset have their lengths ranged from 7 to 25 amino acids (see Fig. 2**c**). So this study zooms into peptides whose lengths ranged within [7,25] with their MIC values against *S. aureus*. The consolidated benchmark dataset consists of 2758 AMPs in total.

Landscape of the benchmark dataset. In order to figure out the challenges in the prediction of MIC values, we take the last layer of the pre-trained 33-layer ESM2 model as the representation of peptides, and employ t-SNE to visualize the two-dimensional distributions of the AMPs. Then we visualize the negative log10 of the MIC values as the z-axis of the overall landscape of the benchmark dataset. The unsmooth landscape intuitively illustrates the existences of many AC pairs (Fig. 2**e**). Fig. 2**f** shows 5 pairs of potential ACs and the complete list of the defined AMPCliff pairs is available in Supplementary Table S6 and S7.

Study rationale and experiment design

How to define activity cliff of AMPs? Here we take two AMPs ALWKTLLKKVLKA as *sequence1* and ALWKTLLKKVLKAAA as *sequence2* to illustrate the procedure (Fig. 3). AMPCliff firstly aligns the two peptides using the Smith-Waterman algorithm. A substitution score matrix is used to calculate the similarity score between the two aligned peptides, and the residue-wise

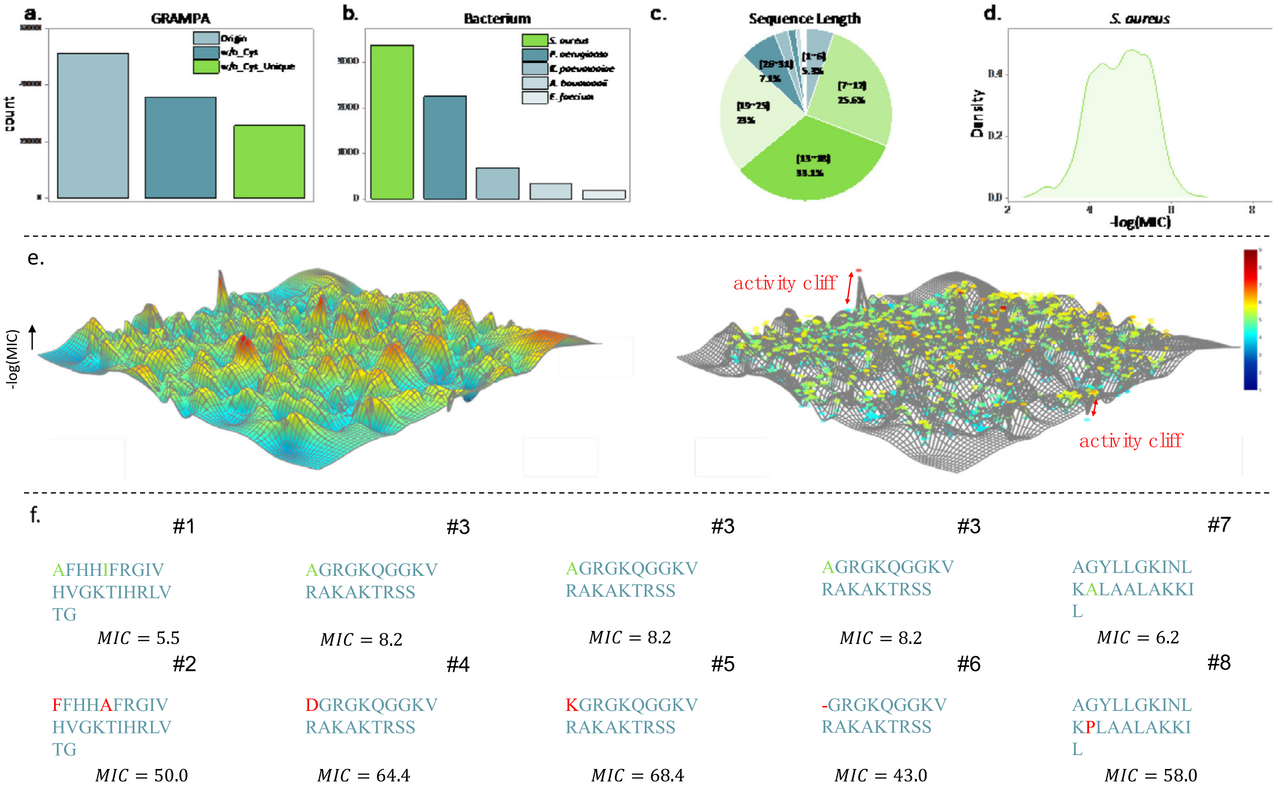


Fig. 2. Exploratory profiling for GRAMPA dataset. a The number of filtered peptides by excluding sequences with Cysteine (w/o_Cys) and dropping duplicate sequences (w/o_Cys_Unique) from the original GRAMPA dataset (Origin). **b** The number of peptides against a bacterium after filtering. *S. aureus, Pseudomonas aeruginosa* (*P. aeruginosa*), *Klebsiella pneumoniae* (*K. pneumoniae*), *Acinetobacter baumannii* (*A. baumannii*), *Enterococcus faecium* (*E. faecium*) are the 5 out of the 6 pathogens that pose the greatest threat to human life. **c** Length distribution of peptides against *S. aureus*. **d** Label distribution of peptides against *S. aureus*. **e** The landscape of the peptides against *S. aureus*. Features were the last layer of ESM2 with 33 layers reduced by t-SNE with two dimensions. **f** ACs showcased on a series of peptides with 5-fold MIC changes. The unit of MIC is μM, and the original data are available in the GRAMPA dataset.

geometric mean is calculated as the final similarity score between these two peptides. Fig. 3**g** shows the MIC difference of *sequence1* and *sequence2* is around 11-fold, large enough to be treated as a pair of AMPCliff.

This study defines a pair of AMPs as AMPCliff, if these two AMPs with canonical amino acids are significantly similar to each other while their antimicrobial activity measurements, i.e., MIC values, have a large fold change (Fig. 3**g**). This is different to the definition of MMP-Cliff in [3] that the similarity between two molecules is measured on the atom-level and the potency change has to be at least 2 orders of magnitudes. We believe that the inter-molecular similarity in AMPCliff aligns more closely with the modular structure of amino acids in AMPs. The following sections will provide a detailed definition of AMPCliff.

Does scaling law still work on AC property prediction? The scaling law refers to that the performance of a language model (LM) improves consistently with the continually increased scale of this LM. This pattern was also observed in protein language models (language models trained by protein data) [41,42]. Therefore, this naturally raises a question: Does the scaling law of GLMs still hold effective in predicting the AC property in the AMPs with canonical amino acids? This question motivates us to evaluate the performance of transformer encoder- and transformer decoder-based models in increasing model sizes (illustrated in Fig. 3**c**). Moreover, some researchers have found that traditional machine learning methods outperformed advanced deep learning methods on AC property prediction tasks [1]. So we choose tree-based models, GP, and SVM as baselines (Fig. 3**a**). We also evaluate the effectiveness of published regression models for peptide design, including Huang et al.'s work [18], Pandi et al.'s work [19], and Schissel et al.'s work [43] Fig. 3**b**). The details of the experimental settings of this work can be found in Supplementary Table S12. If the corresponding methods were derived from the literature, the column "citation" will list its abbreviation in this paper and give the original reference. The hyperparameters of each model are listed in Supplementary Table S1.

How to evaluate the model performance on AC property prediction? The dataset splitting strategy is essential to develop a stable model [44]. ACNet [11] utilized random split and target split to evaluate the model performance. But random split may separate a pair of AMPs one AMPCliff into training and test sets. This separation will reduce the influence of the ACs on the model performance and overestimate the capacity of the model on AC prediction tasks. While target split may introduce an out-of-distribution (OOD) issue into the model and amplifies the influence of ACs on the model performance.

This study introduces AC split, a data split strategy to ensure that AMPCliff pairs appear exclusively in the test set, and evaluates whether a peptide representation method can effectively learn the high-frequency patterns commonly found in AMPCliffs from the low-frequency patterns in training set. The AC split strategy (Fig. 3**f**) firstly extracts all AMPCliff pairs with k-fold changes in the MIC values from the given dataset to form the test set. The remaining peptides are then divided into training and validation sets using a stratified random split strategy, consistent with the approach described in the literature [45]. Further discussion of AC split with other data split strategies can be found in Supplementary Material S2.

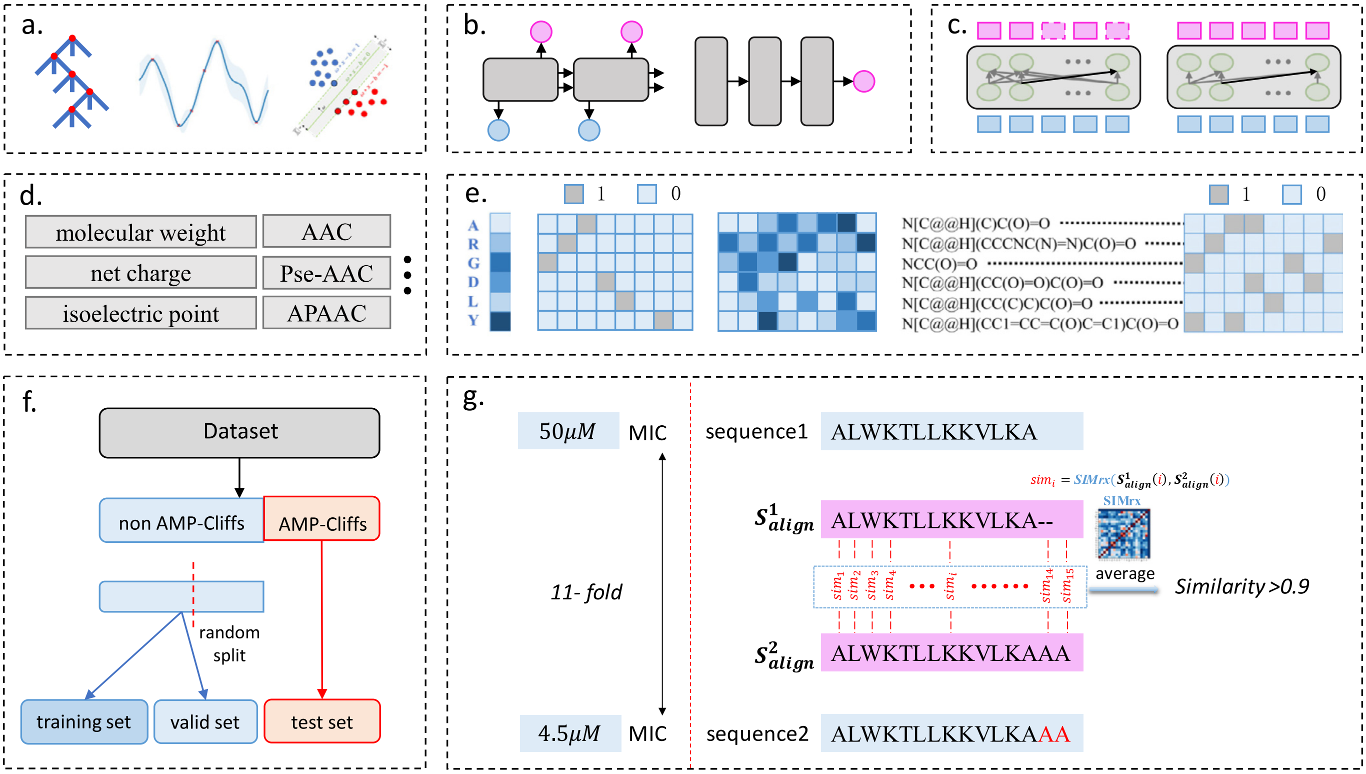


Fig. 3. An overview of the quantitative definition and benchmarking of AMPCliff. a-c are the model architectures used in this paper, including **a** machine learning methods: RF, XGBoost, GB, GP and SVM. **b** deep learning methods LSTM and CNN. **c** pre-trained language models like transformer encoder-based models BERT, ESM2 and transformer decoder-based model GPT2, ProGen2. **d-e** list the features used in this paper, including **d** fixed-type representations, **e** dictionary index, one-hot encoding, word embedding and fingerprint. **f** the procedure of AC split. **g** an illustrative definition of AMPCliff.

Results and discussion

AMPCliffs by conserved mutations

We defined a pair of amino acids with a positive BLOSUM62 substitution score as the conserved substitution, and found out that even conserved substitutions can cause AC in AMPs. We illustrated 10 such AMP pairs as examples in Table 1. The complete list of AMPCliffs was released in Supplementary Table S6. For example, the pair of AMPs FLPIIGKLLSGLL and FLPIVGKLLSGLL only had a substitution from I to V (BLOSUM62 score 3) on the 5th residue, while their MIC values were increased from 4.45 µM to 59 µM, with 13.2681-fold changes in the antimicrobial activities. Another pair of AMPs RRWWRWWR (MIC = 1.80 µM) and RRWYRWWR (MIC = 42 µM) carried a conserved substitution from W to Y and their antimicrobial activities received a 23.3116-fold change. The underlying causes of AC are often complex. According to study [47], both KQRWLWLW and KQKWLWLW are cyclic peptides. However, substituting the basic residue arginine (R) with another basic residue lysine (K) leads to a marked decrease in activity against *S. aureus*. The study [47] further reports that KQRWLWLW can self-assemble into nanotubes within synthetic lipid membranes, whereas there is no evidence indicating that KQKWLWLW exhibits similar self-assembly behavior.

In another work [48], a series of structurally modified synthetic parasin I analogs were analyzed. The structure of parasin I in 30 mM SDS was determined to include a short α-helix (residues 9–17) flanked by disordered regions (residues 1–8 and 18–19). The results demonstrated that a basic residue at the N-terminus is critical for membrane-binding activity. Moreover, the results showed that even a single lysine residue located in a random coil region can profoundly influence the peptide’s action mechanism.

Table 1
Ten example pairs of AMPCliffs with conserved mutations. The two AMPs were aligned by the Smith-Waterman alignment with BLOSUM62 substitution matrix and set penalties of gap open/extension −11 and −1, which were default parameters in BLASTp [46]. The “Mutation” column indicates “(position in the alignment, amino acid in AMP1, amino acid in AMP2)”. The detailed information can be found in the Supplementary Table S6 and S7.

AMP1	MIC1	AMP2	MIC2	Mutation
FLPIIGKLLSGLL	4.45E-06	FLPIVGKLLSGLL	5.9E-05	(5, I, V)
GLLKKIKWLL	2.84E-05	GLLKRIKWLL	3.23E-06	(5, K, R)
GWLDVAKKIGKAAFNVAKNFI	2.23E-05	GWLDVAKKIGKAAFNVAKNFL	1.7E-06	(21, I, L)
ILPWKWPWWKWRR	2.58E-06	LLPWKWPWWKWRR	1.61E-05	(1, I, L)
INLKAIAALAKKLL	2.15E-05	INLKAIAAMAKKLL	1E-06	(9, L, M)
KKKWLWLW	6.74E-05	KRKWLWLW	8.23E-06	(2, K, R)
KQKWLWLW	3.79E-05	KQRWLWLW	3.7E-06	(3, K, R)
RRLFRRILRWL	5.5E-07	RRLFRRILRYL	4.5E-06	(10, W, Y)
RRWWRWWR	1.8E-06	RRWYRWWR	4.2E-05	(4, W, Y)
RWWRWWR	1.4E-05	RWWRYWR	2.07E-06	(5, W, Y)

Measurement of similarity between two AMPs

Limitation of the MMP-Cliff definition in AMPs with canonical amino acids. During our experimental investigations, we identified a limitation in the current definition of MMP-Cliff. This definition does not account for the structural similarity of substructures that transform within a pair of molecules, nor does it consider the evolutionary context of amino acids. Typically, substructures with structural similarities are likely to exhibit similar functional characteristics. Therefore, a revised definition of activity cliffs (ACs) specific to AMPs with canonical amino acids is necessary.

How to measure the similarity between two AMPs beyond MMP-Cliff? This study investigated five commonly used similarity measurements between two peptides or two molecules to evaluate their performance, including Levenshtein distance (Levenshtein), Levenshtein distance between aligned sequences (Levenshtein_aligned), sequence_identity, an average of Tanimoto similarity between aligned sequences (Tanimoto_average), and an average of BLOSUM62 after normalized between aligned sequences (BLOSUM62_average). Two peptide sequences are aligned by the Smith-Waterman algorithm using biopython version 1.79 with Python version 3.8.18. For more details on why we designed BLOSUM62_average and Tanimoto_average to compare with sequence_identity, please see Supplementary Material S3.

For a pair of sequences *sequence1, sequence2*, we denoted them as S_1, S_2 for short. Then we use sequence alignment algorithm to get the aligned sequence pair, denoted as S^1_{align} and S^2_{align} with the same aligned sequence length L_a. Furthermore, we used a similarity matrix $\boldsymbol{SIMrx} \in \boldsymbol{R}^{N\times N}$ to measure the similarity between two amino acids. For any two arbitrary amino acids p and q, we have their similarity value $\boldsymbol{SIMrx}(p,q) \in [0,1]$, and then we get the similarity between the two sequences $\frac{1}{L_a}\sum_{i=1}^{L_a}\boldsymbol{SIMrx}(S^1_{align}(i), S^2_{align}(i))$.

BLOSUM62 versus Tanimoto similarity on AMPCliff. To compare the BLOSUM62 substitution matrix (with integer values ranging from –4 to 11) and the Tanimoto similarity of each amino acid pair based on ECFP (with float values ranging from 0 to 1) on the same scale, we employ a normalization strategy akin to the MaxMinScaler in the Python sklearn package, with an additional step to ensure the symmetry of the formula:

$$M'_{\mathrm{new}} = \frac{M - \min(M)}{\max(M) - \min(M)} \# \tag{1}$$

$$M_{\mathrm{new}} = \frac{M'_{\mathrm{new}} + {M'_{\mathrm{new}}}^{\mathrm{T}}}{2} \# \tag{2}$$

where M represents the original scoring matrix, and M_{new} is the normalized version.

Fig. 4**a** illustrates the details of the normalized BLOSUM62 matrix (referred to as "BLOSUM62 average") and Tanimoto similarity matrix. We observe that certain amino acids exhibit similar patterns across both matrices. For example, the normalized BLOSUM62 values for phenylalanine (F) with tryptophan (W) and tyrosine (Y) are 0.42 and 0.65, respectively, with phenylalanine (F) being most similar to tyrosine (Y) compared to other amino acids. This similarity is mirrored in the Tanimoto similarity matrix, where the similarity values for phenylalanine (F) with tryptophan (W) and tyrosine (Y) are 0.5 and 0.7, respectively, again showing phenylalanine (F) as most similar to tyrosine (Y).

However, significant differences also emerged. For instance, the normalized BLOSUM62 values between cysteine (C) and arginine (R), asparagine (N), and aspartic acid (D) are all below 0.1. In contrast, the Tanimoto similarities between cysteine (C) and arginine (R), asparagine (N), and aspartic acid (D) are 0.3, 0.5, and 0.52, respectively. This discrepancy arises because Tanimoto similarity does not account for structural similarity at the atomic level, and assumes that all atomic substitutions have the same impact.

As shown in Fig. 4**a**, cysteine (C) contains sulfur (S) and often forms disulfide bonds, while arginine (R), asparagine (N), and aspartic acid (D) have markedly different side chains and chemical properties. Arginine (R) is basic with an amino group (–NH2), asparagine (N) is amidic with an amide group (–CONH2), and aspartic acid (D) is acidic with a carboxyl group (–COOH). These fundamental differences lead to distinct physicochemical properties that Tanimoto similarity fails to capture. The MMP-Cliff definition suffers from a similar issue to Tanimoto similarity, i.e., treating atom-level differences uniformly.

In summary, we combine the advantages of the two sequence identity calculation methods and propose the "BLOSUM62 average" as a more suitable similarity measure for short-length peptides with canonical amino acids.

Global alignment versus local alignment. This study investigates AMPCliffs in peptides with canonical amino acids, and employs the Smith-Waterman local alignment algorithm with the BLOSUM62 substitution matrix. We set the gap opening and gap extension penalties to –11 and –1, which are the default parameters in BLASTp [46]. For more discussion about global alignment and local alignment, please see Supplementary Material S4.

Measurement of bioactivity changes between two AMPs

Setting the fold-change threshold for bioactivities. The definition of an MMP-Cliff [3] requires that the potency difference between two compounds meeting the structural criteria must be at least two orders of magnitude. However, the measurement of bioactivities in AMPs uses a different metric, minimum inhibitory concentration (MIC), to assess antimicrobial potency. A lower MIC value indicates higher antimicrobial activities.

Mouton et al. [49] pointed out that the ISO 20776–2 standard for MIC reproducibility allows for an acceptable deviation of one dilution from the mode in 95 % of cases, corresponding to a range of at least two 2-fold dilutions. To determine an appropriate fold-change threshold for AMPCliffs, we evaluated how different minimum MIC differences in AMPCliff pairs, ranging from a 5-fold to a higher fold change, affect model performance using the Recall metric (see "Performance evaluation of AMPCliff predictions" for details). For the definition of Recall, please see **Evaluation Metrics**.

In this study, an MIC value with the unit mole (M) is firstly converted to –log(MIC) (the logarithm base 10). We defined the MIC values of any two sequences as MIC_1 and MIC_2, with $\mathrm{MIC}_1 \geq \mathrm{MIC}_2$. The fold change is defined as:

$$\frac{\mathrm{MIC}_1}{\mathrm{MIC}_2} \geq \tau \# \tag{3}$$

$$-\log\left(\frac{\mathrm{MIC}_1}{\mathrm{MIC}_2}\right) = -log MIC_1 - (-\log \mathrm{MIC}_2) \leq -\log(\tau) \# \tag{4}$$

$$|-\log(\mathrm{MIC}_1) - (-\log(\mathrm{MIC}_2))| \geq \log(\tau) \tag{5}$$

In order to define the fold change threshold properly, we set τ = 2, 3, 4, and 5 to establish the dataset for model training and performance comparison with recent state-of-the-art methods for the AMPCliff prediction task. (See **Performance evaluation of AMPCliff predictions**) Note that a larger –log(MIC) value indicates better antimicrobial activities.

Experimental data supporting theoretical analysis. To ensure our methods are applicable in real-world scenarios, we compared the number of detected 5-fold change AMPCliffs across different sequence length regions using various similarity metrics: Levenshtein, Levenshtein_aligned, sequence_identity, BLOSUM62_average, and Tanimoto_average. Fig. 4**b** provides a schematic of these

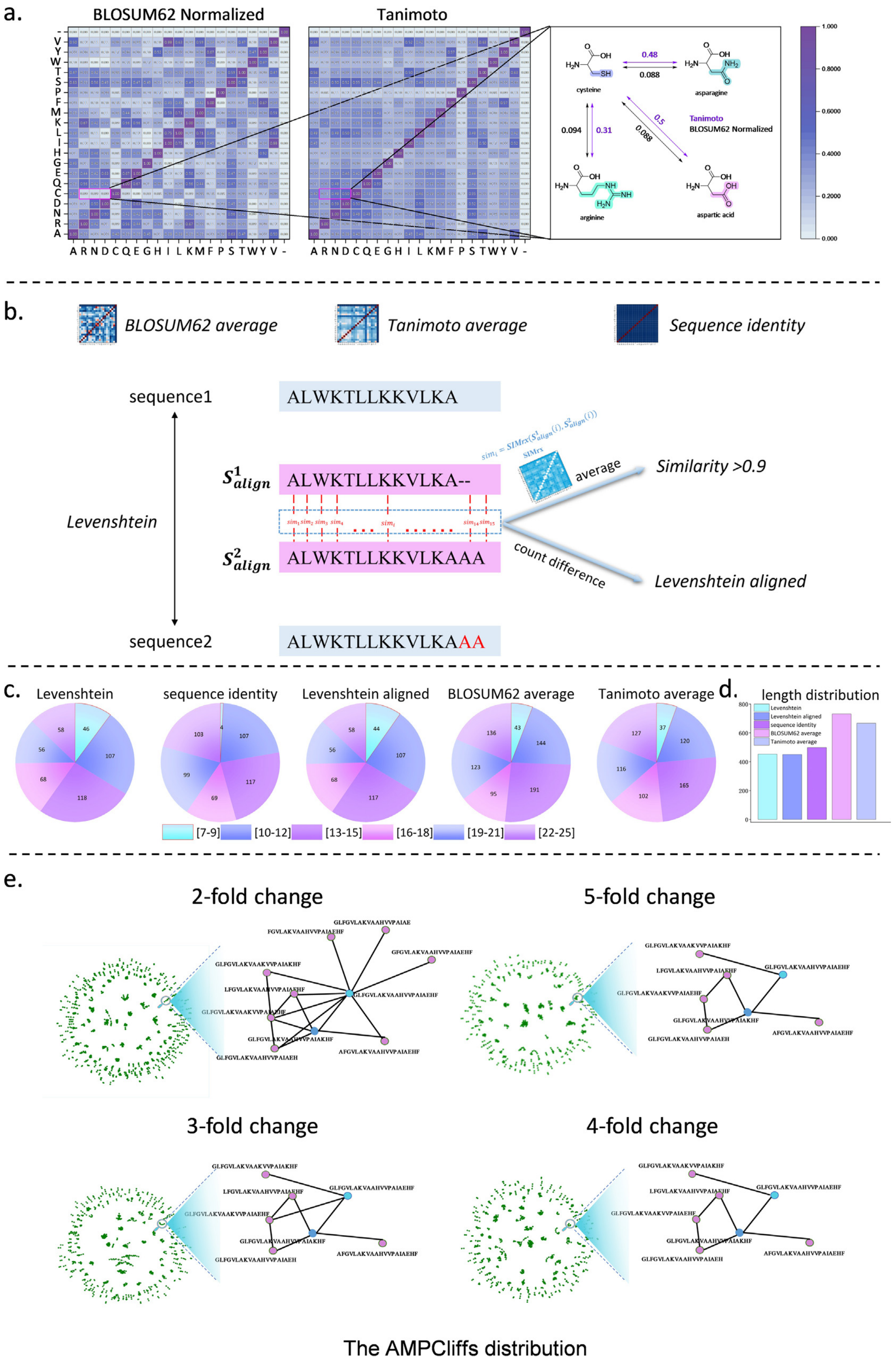


Fig. 4. Similarity measurements between AMPs. a The difference in amino-acid level similarity matrix measured by Tanimoto similarity and BLOSUM62 after normalization. **b** A toy example of 5 similarity measurements, i.e., Levenshtein, Levenshtein_aligned, sequence_identity, Tanimoto_average, and BLOSUM62_average. Levenstein distance measures the minimum number of single-character edits (insertions, deletions, or substitutions) required to change one string into another. In this case, Levenstein distance is 2. sequence_identity, Tanimoto_average, and BLOSUM62_average used the same operation but different similarity metrics. Levenshtein_aligned is to count the difference of the aligned sequence pair. **c** The numbers of AMPCliffs with 5-fold change in MIC values measured by Levenshtein, Levenshtein aligned, sequence identity, BLOSUM62 average, and Tanimoto average. Here we set 1 as the threshold for "Levenshtein" and "Levenshtein aligned", and 0.9 as the threshold for the other 3 methods. **d** the total number of 5-fold change AMPCliffs under 5 similarity measurements. **e** Showcases of the variation of AMPCliffs as the fold change increased. Nodes represented sequences. If two nodes were marked as an AMPCliff pair, then they got an edge. The red node and the blue node represented the top 2 maximum degree nodes as fold change increased. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

measurement procedures, while Fig. 4**c-d** present the numbers of detected 5-fold AMPCliffs across different length ranges.

Levenshtein distance is also known as edit distance, and it measures the difference between two sequences by counting the minimum number of single-character edits (insertions, deletions, or substitutions) required to convert one sequence into the other. The difference between "Levenshtein" and "Levenshtein_aligned" lies in whether the sequences are aligned prior to comparison (see Fig. 4**b**). Our results show that these two metrics yield similar outcomes (see Fig. 4**c**).

The distinction between "Levenshtein_aligned" and "sequence_identity" is that sequence_identity averages similarity over the entire sequence length (see Fig. 4**b**). This averaging operation biases sequence identity toward identifying longer sequences. For example, if two aligned sequences S^1_{align} and S^1_{align} differ by only one residue, the sequence identity will be only 0.9 if the aligned sequences are 10 residues. This bias explains why AMPCliffs identified by sequence identity tend to be longer and why few sequences in the [7–9] length range are detected (see Fig. 4**c**). This also suggests that MMseqs2 is not suitable for clustering short peptides, a conclusion that aligns with experimental findings in the literature [50,51].

The length distributions of AMPCliffs identified by "BLOSUM62_average" and "Tanimoto_average" are similar to each other, with "BLOSUM62_average" generally detecting a high count (see Fig. 4**d**). Each metric has its strengths and limitations. BLOSUM62_average incorporates the prior knowledge of medicinal chemists and evolutionary information, but cannot be extended to the peptides with non-canonical amino acids. Tanimoto_average treats all atomic substitutions equally and disregards evolutionary information. But it can be easily applied to peptides with non-canonical amino acids.

Therefore, this study presents two versions of the AMPCliff identification models. For tasks involving canonical amino acids, the BLOSUM62 version is recommended. For tasks involving non-canonical amino acids, the Tanimoto version is preferred if a suitable sequence alignment algorithm for non-canonical peptides is available. To our knowledge, PepSeA [52] is the only sequence alignment algorithm currently available for such peptides. This paper primarily reports results from AMPCliff predictions using the "BLOSUM62_average" metric, with the detailed results for both "BLOSUM62_average" and "Tanimoto_average" metrics provided in Supplementary Tables S2-S4 and S8-S9.

In summary, our experimental results confirm that sequence identity is not a suitable metric for measuring similarity between short AMPs with canonical amino acids. Instead, our proposed "BLOSUM62_average" measurement can detect the highest number of AMPCliffs.

Quantitative definition of AC in AMPs

We further propose a new quantitative definition of AC in AMPs with canonical amino acids, termed AMPCliff. The principles of AMPCliff align with a common understanding of AC that structurally similar molecules are active against the same tween two AMPs is defi bacteria but exhibit significantly different MIC values. We set a 2-fold change as the minimum threshold for MIC differences. The similarity between the two AMPs is defined as follows:

Given a pair of peptide sequences S_1 and S_2, we first use the Smith-Waterman alignment algorithm to obtain the aligned sequences, denoted as S^1_{align} and S^2_{align}, of the same length, L_a. We utilize a similarity matrix $\boldsymbol{SIMrx} \in \boldsymbol{R}^{N\times N}$ to measure the similarity between two residues in the alignment. For a pair of matched residues p and q in the alignment, their similarity value is defined as $\boldsymbol{SIMrx}(p,q) \in [0,1]$. The overall similarity between the two aligned peptide sequences is then calculated as:

$$\frac{1}{L_a}\sum_{i=1}^{L_a}\boldsymbol{SIMrx}\left(S^1_{align}(i),S^2_{align}(i)\right)\# \qquad (6)$$

Fig. 3**g** illustrates the AMPCliff definition procedure using the two AMP sequences ALWKTLLKKVLKA and ALWKTLLKKVLKAAA as examples.

The key difference between AMPCliff and MMP-Cliff lies in the similarity measurement. AMPCliff first employs sequence alignment to align the two molecules, and then uses a similarity matrix to calculate the similarity of residues at each aligned position. The final similarity is obtained by taking the geometric mean of these values. Further discussion of similarity threshold criteria can be found in Supplementary Material S5. Having established the definition of AC in AMPs with canonical amino acids (AMPCliff), the next question is how we can evaluate model performance in predicting AMPCliffs?

Performance evaluation of AMPCliff predictions

We compare the performance of a 4:1 stratified random split (referred to as "random split") with our proposed AC split, varying the MIC threshold from 2-fold to 5-fold changes. We denote the AC split strategy with the *k*-fold changes in the MIC threshold as *k*-fold AC split, where *k* = 2, 3, 4, or 5. The model does not encounter high-frequency data in the AMPCliffs in the training set when using the AC split strategy, while some sequences in the AMPCliffs may appear in the training set of the random split strategy. The performance of the random split strategy is expected to be overestimated. However, our observations suggest that this common assumption may not always hold true in practice.

As shown in Fig. 5**(a-b)**, the choice of performance metric can significantly influence the perceived model performance. For instance, the difference in the metric SCC (upper panel of Fig. 5**a**) and other metrics like PCC and R2 (Fig. 5**b**) between random split and AC split with 2/3/4/5-fold changes are generally positive (negative for RMSE and MAE). This indicates that random split is indeed overestimated, consistent with the literature [44]. However, the Recall metric (lower panel of Fig. 5**a**) does not follow the same trend. As the fold change increases from 2 to 5, the model's performance with random split initially outperforms the AC split with 2-fold change, then becomes competitive to the AC splits with 3- and 4-fold changes, and eventually underperformed relative to the AC split with 5-fold change. This phenomenon is intriguing because the test set of the AC split with 2-fold change includes AMPCliffs with 2-fold and larger changes, whereas the test set of the AC split with 5-fold change only includes AMPCliffs with MIC changes of 5-fold or greater. This result suggests that while the random split can predict low-frequency (2-fold to 3-fold) and relatively higher-frequency data (3-fold to 4-fold) well, it struggles with very high-frequency data (5-fold and above).

If we change the data split strategy from random split to AC split, the models may perform poorly in predicting exact MIC values (with lower Pearson, R2, RMSE, and MAE compared to random split) and general MIC ranking (lower Spearman correlation). However, they can improve the Recall of the top 50 sequences, which is crucial in practice as we often focus on the top *k* predictions. Consequently, we determined that a 5-fold change is the maximum threshold for defining an AMPCliff.

We further examine the results of the pre-trained language models (LMs) on the 2-fold and 5-fold AC splits across training, validation, and test sets (Fig. 5**c-d**). The validation set performance for the 2-fold AC split is less consistent compared to the 5-fold AC split, which is understandable given that a 2-fold change repre-

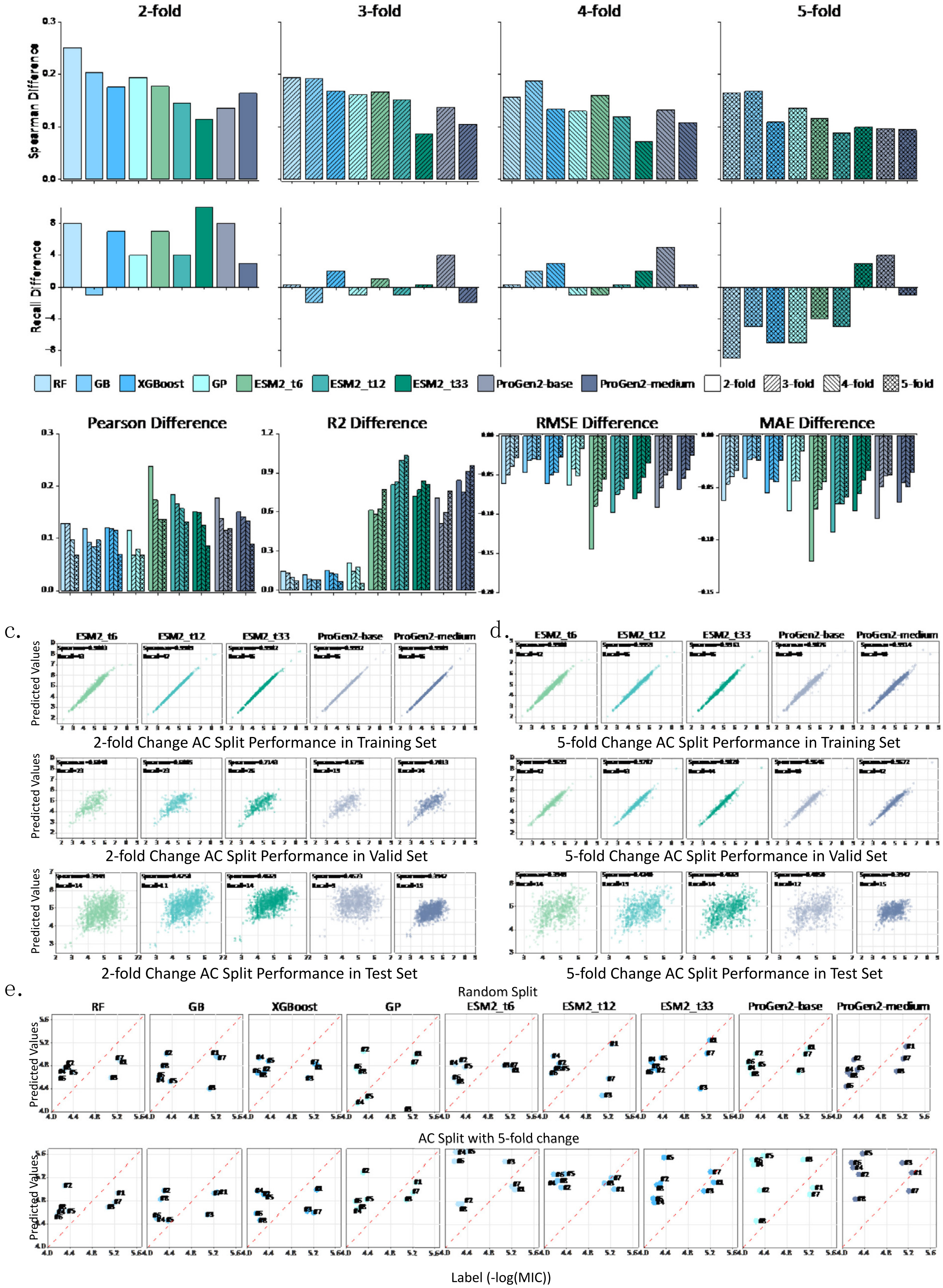


Fig. 5. Performance comparison between random split and AC split methods. Panels **a-b** show the difference in Spearman, Recall **a** and other performance metrics **b** Metric(random split) – Metric(AC split), varying from 2-fold to 5-fold changes in the MIC threshold. Panels **c-d** present scatter plots for the training/validation/test sets under 2-fold **c** and 5-fold **d** in the MIC threshold and AC splits. Panel **e** displays a scatter plot of ACs with 5-fold changes in the MIC threshold from Fig. 2**f**, compared with random split, where the red line represents the diagonal line $y = x$. The label is the –log(MIC) value of each AMP. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

sents the minimum reproducibility tolerance for the same peptide on a MIC detection device [49]. This suggests that the training set for the 2-fold AC split contains more systematic noise. Conversely, the 5-fold AC split results indicate that the LMs are well-trained on peptides within the 5-fold change range. Although these models are poor at predicting ACs, they have still learned some general knowledge from non-AC peptides (low-frequency data) to AC peptides (high-frequency data), supporting the rationale for using a 5-fold threshold in our AMPCliff definition. For further discussion of the Case Studies of AMPCliff Predictions (Fig. 5**e**), please see Supplementary Material S6.

Learned representations versus fixed type representations

The learned representations refer to those peptide features derived from the pre-trained DL models, while the fixed type representations are defined and calculated by manually formulated equations. To check if learned representations outperform fixed type representations, we compare ML models GP and SVM with published deep learning methods AMPSpace, CellFree-rnn, CellFree-cnn and peptimizer on AMPCliffs defined by "BLOSUM62 average" with 2/3/4/5-fold changes. Notice that we add a new model SVM into ML methods for comparison, for checking whether it is suitable for AMPCliff prediction.

Firstly, fixed type representations with ML methods outperform the conventional DL methods. We continue to consider Recall and SCC as the performance metrics since the relative order of an AMPCliff pair's predicted values is more valuable than the values themselves in the drug discovery process. We calculate Recall for the top 50 of the label values and their predicted values (see **Evaluation Metrics** for the definition of Recall). Since we use –log(MIC) as the label value, a larger value means the AMP has better antimicrobial activity. As shown in Fig. 6**a-b,** RF gets the best SCC value with 4-fold and 5-fold changes. XGBoost achieves the best SCC value with 2-fold and 3-fold changes. Meanwhile, RF and GB achieve the best Recall with 5-fold change and 2-fold change, respectively. GP gets the best Recall value with 3-fold and 4-fold changes. SVM performs poorly compared with the other ML methods, and none of the four DL methods outperform the best traditional methods (namely, the fixed type representations with ML methods.), which is consistent with the literature [31]. All of the best performances by AC split across 2/3/4/5-fold changes belong to the fixed type representations with ML methods.

Moreover, it turns out that the fixed-type representations with ML methods can transfer some general patterns from low-frequency data to high-frequency data. Table 2 shows the Recall of the fixed type representations with each ML method by the AC Split with 2/3/4/5-fold changes, namely gradually adding more low-frequency data to the training data. Except for XGBoost, the Recall of all the other ML methods increases or stays at the same value as the training data seeing more low-frequency data. Besides, the XGBoost gets 9 for 4-fold change and 10 for 3-fold change, which is quite close to each other. More evidence may be found in the Supplementary Table S2.

The learned representations based on LMs, especially ESM2, have the potential to beat the fixed-type representations. Since SVM is the poorest model, we remove it from further analysis. Then we compare the four ML methods, RF, GB, XGBoost, and GP with 4 MLMs: BERT bae version, ESM2 with 6 layers, ESM2 with 12 layers, and ESM2 with 33 layers. As shown in Fig. 6**e-f,** the largest model ESM2 with 33 layers beats the ML methods and achieves the best performances in SCC among all fold changes. Besides, there is an interesting ascending trend in the metric SCC of ESM2 as the model size gets larger. In the meantime, in terms of Recall, MLMs get the best Recall compared with the ML methods in general, but ESM2 doesn't have the same ascending trend as it does in SCC. Another noteworthy result is that the prediction values of the English text-based pre-trained model BERT perform similarly to ESM2 with 12 layers. Although its SCC is the worst among all the MLMs, its Recall metric is competitive with ESM2. Fig. 6**i** calculates the statistical significance of the null hypothesis of the predicted values of two models are the same by the Mann-Whitney *U* test. The p-value smaller than 0.05 means that the predicted values of the compared models is significantly different. It turns out that the predicted values of BERT with the base version don't have a significant difference compared with ESM2 with 12 layers among all fold changes.

Meanwhile, we compare the four ML methods with four GLMs: GPT2, Progen2 small version, base version, and large version. As shown in Fig. 6**g-h,** it turns out that GLMs perform worse than ML methods in SCC and only the Recall of the medium size of ProGen2 beats ML methods on the 2-fold and 3-fold changes, and the second-best Recall on the 2-fold and 3-fold change are still the ML methods GB and GP.

ESM2 with 33 layers beats the fixed types representations in SCC and the learned representations based on MLMs can beat the fixed type representations in Recall, while ProGen2 or the conventional DL methods remain to be further improved. We speculate that the MLM-based learned representations can be good solutions for solving the AMPCliff prediction task. This training paradigm can capture some structural information from peptide sequences, as claimed in ESM2 [41] and ESM3 [53].

Does scaling law still work?

Here we take the scaling laws defined by OpenAI into consideration [54]. Model performance heavily depends on scale, which consists of three factors: the number of model parameters (excluding embeddings), the size of the dataset, and the amount of computing power used for training [54]. Since we utilize the pre-trained language models to fine-tune the AMPCliff prediction task, and evaluate whether there is an ascending trend on each performance metric as the model scale gets larger.

To study whether the scaling laws still work on the AMPCliff prediction task, we evaluate the prediction performance of ESM2 with 6/12/33 layers and ProGen2 small/base/medium versions under the "BLOSUM62 average" condition by AC split. The analysis aims to answer the question: does scaling law still work on the AMPCliff prediction task? Note that BERT and GPT are removed from this analysis because they are not in the ESM2 or ProGen2 series.

Firstly, we investigate their performance in the metrics SCC and Recall, and find out that ESM2 has an ascending trend in predicting the relative order of AMPCliffs. As shown in Fig. 7**a**, there is an obvious SCC improvement as the model size of ESM2 increases. This trend exists almost in all fold changes. However, it disappears in the Recall metric (see Fig. 7**b**), which suggests that the top 50 samples in the test dataset don't perform better as the model size of ESM2 increases. At the same time, GLMs don't show a similar trend as MLMs, i.e., there is neither a significant increasing trend in SCC nor Recall except for the 2- and 3-fold changes. This suggests that GLMs are possibly good at predicting the top 50 samples in AMPCliffs other than the general performance. Besides, if we compare the other performance metrics as shown in Fig. 7**c**, it turns out that both in MLMs and GLMs, the trend of PCC is different from SCC which is extremely high in 2-fold change whereas R2, RMSE, and MAE are getting worse as the model size increases. As the fold change reaches to 5-fold change, the significance of the predicted values among the language models (both MLMs and GLMs) become mostly the same ($p > 0.05$, see Fig. 7**d**). The results indicate that although pre-trained language models fail to predict

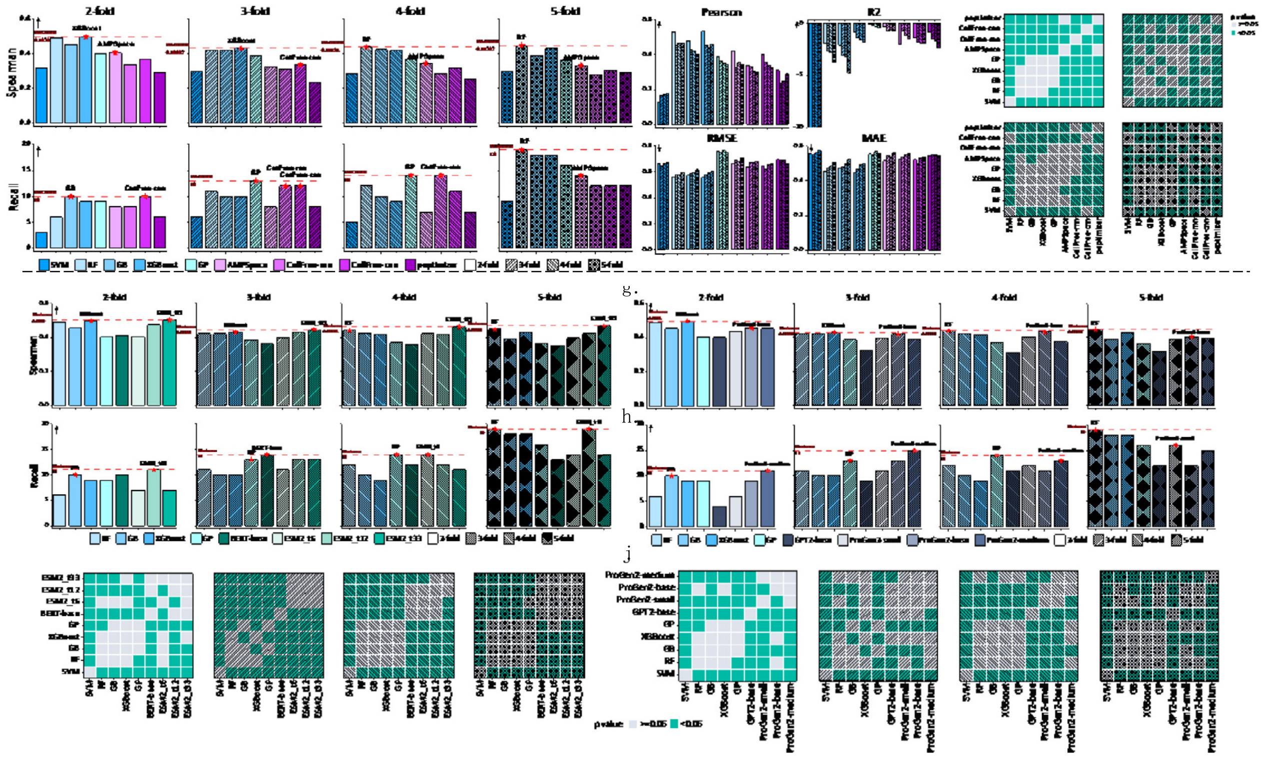


Fig. 6. Prediction performance with the benchmark dataset by AC split. The performance metrics **a** Spearman correlation coefficient and **b** Recall of five ML methods SVM, RF, GB, XGBoost, and GP on fixed type representations and four DL methods AMPSpace, CellFree-rnn, CellFree-cnn and peptimizer on the AMPCliffs defined by "BLOSUM62 average" with 2-fold, 3-fold, 4-fold and 5-fold change. **c** Other prediction performance metrics of the models. **d** Statistically predicted value significance for pairwise model comparison by Mann-Whitney *U* test. **e** Spearman and **f** Recall of five ML methods SVM, RF, GB, XGBoost, and GP on fixed representations and four masked language model methods BERT, ESM2_t6, ESM2_t12 and ESM2_t33 on the AMPCliffs with 2-fold, 3-fold, 4-fold and 5-fold change. **g** Spearman correlation coefficient and **h** Recall of five ML methods SVM, RF, GB, XGBoost, and GP on fixed representations and four generative language model methods GPT, ProGen2-small, ProGen2-base and ProGen2-medium on the AMPCliffs with 2/3/4/5-fold changes. Statistically predicted value significance for pairwise model comparison by Mann-Whitney *U* test on **i** the four masked language models (MLMs) BERT, ESM2_t6, ESM2_t12 and ESM2_t33 and **j** four generative language models (GLMs) GPT, ProGen2-small, ProGen2-base and ProGen2-medium.

Table 2
The metric Recall of the fixed type representations with each ML method by the AC Split with 2/3/4/5-fold changes. The columns are the AC Split with 2-fold to 5-fold change, and a row represents each ML method. The ML methods use the same fixed-type representation features (see **Evaluation Metrics** for the definition of Recall).

Fold	2	3	4	5
SVM	3	6	5	9
LR	9	9	10	12
L1	9	8	11	10
L2	8	8	8	15
ElasticNet	7	8	8	10
RF	6	11	12	19
GB	10	10	10	18
XGBoost	9	10	9	18
GP	9	13	14	16

high-frequency data accurately in the AMPCliff framework, ESM2 has a potential scaling law of predicting AC well.

To investigate the capability of these LMs in general, we calculate the average values of the performance metrics of each MLM or GLM among all fold changes. As shown in Fig. 7**e**, it turns out that except for Recall and R2, all of the other 4 metrics get better as the model size of ESM2 increases. Whereas no metrics had the similar trend on GLMs (Fig. 7**f**). Note that all the R2 values are negative, indicating that the models are poor at predicting the label value −log(MIC) accurately.

We conclude the scaling law of ESM2 is available on the AMPCliff prediction task, while ProGen2 doesn't.

Conclusions

This study identified and systematically defined the activity cliff (AC) phenomenon in AMPs (AMPCliff) as a variant of AC, and captured structurally similar peptides with significant differences in MIC values of antimicrobial activities. We leveraged the GRAMPA dataset to illustrate AMPCliffs with *S. aureus* as a reference, and conducted a comprehensive benchmark analysis of predictive algorithms in this context. Besides an additional α-helix structural descriptor, the total list of the other 676 fixed-type representation features can be found in [18]. In general, any descriptors that capture the physicochemical and structural properties of AMPs can be considered. Additionally, learned representations from large-scale pre-trained models (e.g., the ESM family) are also applicable and have been evaluated in our benchmarking framework.

Accurately quantifying peptide similarity proved essential, as existing sequence identity measures inadequately addressed short peptides. By developing a similarity calculation that considers the average alignment column similarity, we reframed peptide similarity as an amino-acid-level comparison. Our analysis of the BLOSUM62 substitution matrix is enriched with structural insights over traditional metrics like Tanimoto similarity, and underscores the need for a subtle similarity measure for non-canonical substructures. Additionally, our results suggest that current amino-acid-level models may benefit from greater structural detail within amino acid side chains, potentially enhancing models like ESM2 for future AMP prediction tasks.

We also introduced the AC split strategy, and its comparison against random split revealed that models trained on AC split per-

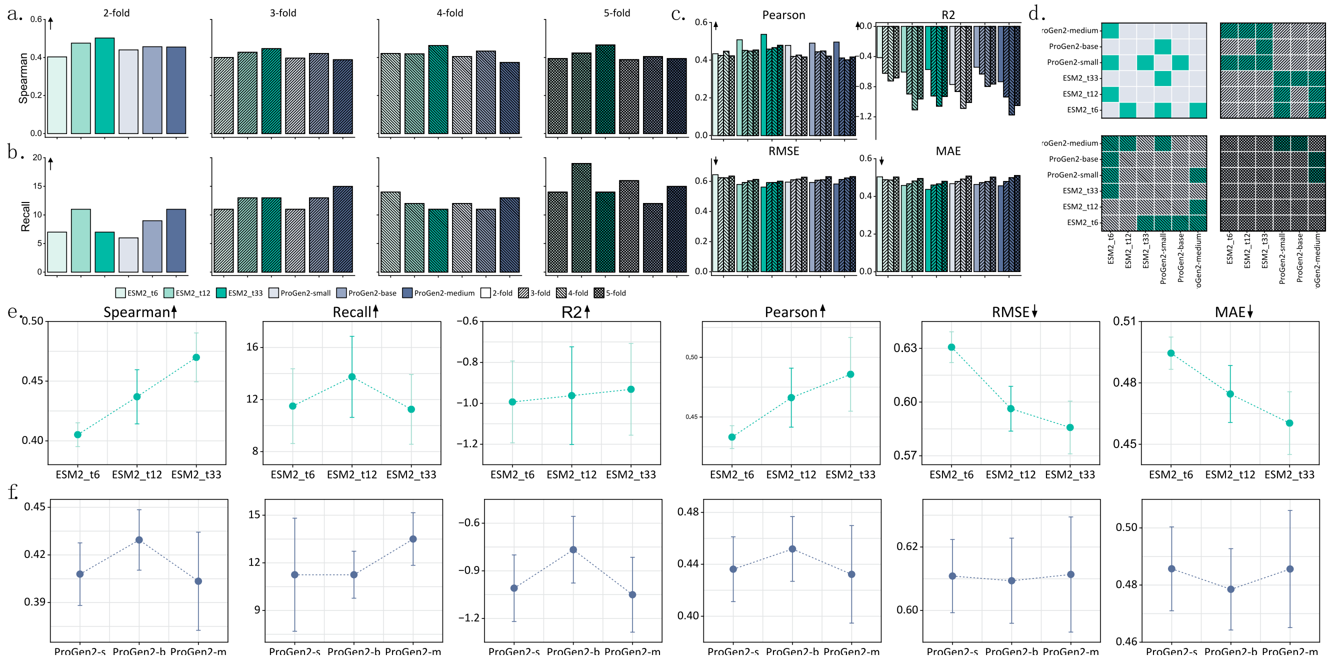


Fig. 7. Evaluating the prediction performance of MLMs and GLMs. **a** SCC and **b** Recall of three MLMs and three GLMs on the AMPCliffs defined by "BLOSUM62 average" with 2–5 fold changes. **c** Other prediction performance metrics of the models. **d** Statistical predicted value significance for pairwise model comparison by Mann-Whitney *U* test. The average performance of each **e** MLM or **f** GLM with 2–5 fold changes. Note that "ProGen2-s" is short for "ProGen2-small", "ProGen2-b" is short for "ProGen2-base", "ProGen2-m" is short for "ProGen2-medium".

formed better on detecting AMPCliffs with high-frequency signals based on the transferable knowledge learned from non-AC peptides. Our evaluation metrics, such as Recall and SCC, emphasized the predictive capacity across various ML and DL models. ESM2 demonstrated superior performance and scalability in AMPCliff prediction, and outperformed both fixed type and GLM-based learned representations, particularly with larger model sizes.

Our findings revealed limitations in current deep learning-based representation models. To better capture the properties of antimicrobial peptides (AMPs), it is necessary to incorporate atomic-level dynamic information related to their mechanisms of action. For instance, AMPs that disrupt cell membranes may self-assemble into polymeric structures that interact with membranes in complex ways [55], which cannot be fully represented by sequence information alone. Therefore, integrating molecular interaction trajectories between AMPs and membranes into the modeling process represents an important direction for future research.

Limitations

This study explores how to better define activity cliffs in peptides composed of canonical amino acids. When encountering non-natural amino acids, although the Tanimoto Similarity is less rigorous than BOLSUM62, we have yet to find a better matrix that describes the similarity between descriptor structures. Additionally, as the only publicly available peptide dataset with MIC values is GRAMPA, the datasets we can construct are relatively small. Therefore, our AMPCliff search algorithm is primarily designed for small datasets, as it becomes computationally time-consuming for larger datasets (exceeding 5,000 entries).

Besides, when the dataset is small, the deep learning algorithm might easily encounter overfitting problems by training. In this case we recommend that practitioners use the pre-trained model weights directly for prediction, with minimal or no fine-tuning, and instead incorporate feature selection and traditional machine learning methods for modeling, as suggested in the literature. Moreover, when both the training and test sets are small in size, selecting appropriate performance metrics becomes critical, as limited data can cause high variance in metric values. For example, in an extreme case where the test set contains only two samples with very similar MIC values, the Spearman correlation may fluctuate drastically between −1 and 1.

Finally, our model results are mostly based on fine-tuning the pre-trained models on antimicrobial peptides (AMPs) against *S. aureus*. They can be transferred to the close organisms like *Staphylococcus epidermidis* (*S. epidermidis*), but we recommend fine-tuning on new organisms, especially for more distantly related species such as *E. coli*. We showed the inference results of *S. epidermidis* and *E. coli* by the ESM2 with 6 layers model trained by *S. aureus* in Supplementary Material S7.

Materials and methods

Datasets

In 2018, Jacob Witten and Zack Witten introduced the GRAMPA dataset [39], which remains one of the only AMP datasets containing minimum inhibitory concentration (MIC) values. Previous studies have utilized GRAMPA for developing machine learning and deep learning models in AMP design [18,19], such as AMP-Space [18], CellFree-cnn [19], and CellFree-rnn [19], which we also employ here for comparative analysis on the AMPCliff prediction task.

To evaluate model performance, we implemented both random split and our proposed AC split strategies. Traditional data splitting methods commonly applied in small molecule drug design – such as scaffold, stratified, time, and target splits – are adapted to fit specific data characteristics. For example, time split captures historical trends in drug optimization [56], and scaffold split approximates these temporal patterns [57]. Target split [11] clusters data based on a shared target, fostering deep model learning of structural similarities, while avoiding out-of-distribution (OOD) issues associated with sequence diversity [11]. However, these methods are not ideally suited for antimicrobial peptide datasets like GRAMPA, which lacks time information and is highly heterogeneous, containing a variety of modified peptide structures absent in controlled corporate datasets.

Given these limitations, we developed the AC split method, inspired by previous work by Deng et al. [44] and Dablander et al. [1], which facilitates model learning from non-AC (low-frequency) to AC (high-frequency) peptides. In our implementation, AMPCliff pairs are progressively introduced to the training set by adjusting the fold-change threshold from 2 to 5, incrementally adding high-frequency data to evaluate model adaptability to AMPCliff predictions.

For comprehensive assessment, we employed both the AC split and a 5-fold stratified random split.

Evaluation metrics

Model evaluation procedure can be substantially influenced by the choices of statistical analysis, metrics, and task settings [44]. We selected Spearman's rank correlation coefficient (SCC) and Recall as our primary evaluation metrics, prioritizing the relative ranking accuracy of predictions over their absolute label –log (MIC) values. Additional metrics include R squared (R2), Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Pearson's correlation coefficient (PCC). Their detailed definitions may be found in Supplementary Material S1.

Model training

This study used default hyperparameter values for ML models. For DL models, we followed hyperparameter values reported in the literature. Specifically, following [18], we configured the LSTM with a hidden dimension of 128, embedding dimension of 50, dropout of 0.7, 2 layers, and unidirectional mode. For CNN and RNN models, we set the CNN kernel size to 5, with 64 output channels across 2 layers, followed by 3 fully connected layers [19]. The RNN (LSTM) model used a hidden dimension of 500 and was followed by a single fully connected layer. Following [43], we set the fingerprint radius to 3 and bit count to 2048, with a CNN of kernel size 2, 256 output channels across 3 layers, and a final fully connected layer. Detailed configurations are in Supplementary Table S1. We retained the default hyperparameters for language models as outlined in Supplementary Table S5.

All experiments with neural network and language models were conducted on a single NVIDIA V100 GPU for 100 epochs and 50 epochs, respectively. The validation loss guided the selection of the best model during training for final testing. Batch size was set to 4, as smaller batches facilitate learning of the refined, localized features characteristic of AC, while larger batches typically aid in capturing broader and general features. To ensure a fair comparison, we saved all raw predictions, and evaluation metrics were consistently computed using identical code.

Statistical analysis

To assess statistically significant differences among models and representations, we employed methods consistent with prior work (Deng et al., 2023), and conducted comprehensive statistical analyses on prediction performance (Supplementary Tables 9–11). Specifically, we utilized the non-parametric Wilcoxon rank-sum test to account for non-normal data distributions. A significance threshold of two-sided $p < 0.05$ was applied throughout.

Code and data availability

Resources and benchmark code are available at https://www.healthinformaticslab.org/supp/ or https://github.com/Kewei2023/AMPCliff.

CRediT authorship contribution statement

Kewei Li: Conceptualization, Methodology, Software, Data curation, Writing – original draft, Writing – review & editing. **Yuqian Wu:** Visualization, Software. **Yinheng Li:** Visualization, Writing – review & editing. **Yutong Guo:** Visualization. **Yanwen Kong:** Visualization. **Yan Wang:** Data curation. **Yiyang Liang:** Data curation. **Yusi Fan:** Visualization. **Lan Huang:** Funding acquisition. **Ruochi Zhang:** Conceptualization, Methodology, Project administration. **Fengfeng Zhou:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work was supported by the Senior and Junior Technological Innovation Team (20210509055RQ), the National Natural Science Foundation of China (62072212 and U19A2061), Guizhou Provincial Science and Technology Projects (ZK2023-297), the Science and Technology Foundation of Health Commission of Guizhou Province (gzwkj2023-565), Science and Technology Project of Education Department of Jilin Province (JJKH20220245KJ and JJKH20220226SK), the Jilin Provincial Key Laboratory of Big Data Intelligent Computing (20180622002JC), the Development Project of Jilin Province of China (No. 20220508125RC), and the Fundamental Research Funds for the Central Universities, JLU. We sincerely thank the anonymous reviewers for their valuable time, constructive feedback, and insightful suggestions, which have greatly improved the quality and clarity of this manuscript.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.jare.2025.04.046.

References

[1] Dablander M, Hanser T, Lambiotte R, Morris GM. Exploring QSAR models for activity-cliff prediction. J Cheminform 2023;15:47.

[2] Guha R, Van Drie JH. Structure–activity landscape index: identifying and quantifying activity cliffs. J Chem Inf Model 2008;48:646–58.

[3] Hu X, Hu Y, Vogt M, Stumpfe D, Bajorath J. MMP-Cliffs: systematic identification of activity cliffs on the basis of matched molecular pairs. J Chem Inf Model 2012;52:1138–45.

[4] Hu Y, Bajorath J. Exploration of 3D activity cliffs on the basis of compound binding modes and comparison of 2D and 3D cliffs. J Chem Inf Model 2012;52:670–7.

[5] Hu Y, Furtmann N, Gütschow M, Bajorath J. Systematic identification and classification of three-dimensional activity cliffs. J Chem Inf Model 2012;52:1490–8.

[6] Muratov EN et al. QSAR without borders. Chem Soc Rev 2020;49:3525–64.

[7] Seebeck B, Wagener M, Rarey M. From activity cliffs to target-specific scoring models and pharmacophore hypotheses. ChemMedChem 2011;6:1630–9.

[8] Stumpfe D, Hu H, Bajorath J. Evolving concept of activity cliffs. ACS Omega 2019;4:14360–8.

[9] Stumpfe D, Hu H, Bajorath J. Advances in exploring activity cliffs. J Comput Aided Mol Des 2020;34:929–42.

[10] van Tilborg D, Alenicheva A, Grisoni F. Exposing the limitations of molecular machine learning with activity cliffs. J Chem Inf Model 2022;62:5938–51.

[11] Zhang Z, Zhao B, Xie A, Bian Y, Zhou S. Activity Cliff Prediction: Dataset and Benchmark. arXiv preprint arXiv:2302.07541 (2023).

[12] Silipo C, Vittoria A. QSAR, rational approaches to the design of bioactive compounds; 1991.
[13] Sisay MT, Peltason L, Bajorath J. Structural interpretation of activity cliffs revealed by systematic analysis of structure–activity relationships in analog series. J Chem Inf Model 2009;49:2179–89.
[14] Borkar MR, Pissurlenkar RR, Coutinho EC. Mapping activity elements of protegrin antimicrobial peptides by HomoSAR. RSC Adv 2015;5:78790–8.
[15] Wang L et al. Therapeutic peptides: current applications and future directions. Signal Transduct Target Ther 2022;7:48.
[16] Tropsha A, Isayev O, Varnek A, Schneider G, Cherkasov A. Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR. Nat Rev Drug Discov 2023.
[17] Sadybekov AV, Katritch V. Computational approaches streamlining drug discovery. Nature 2023;616:673–85.
[18] Huang J et al. Identification of potent antimicrobial peptides via a machine-learning pipeline that mines the entire space of peptide sequences. Nat Biomed Eng 2023;7:797–810.
[19] Pandi A et al. Cell-free biosynthesis combined with deep learning accelerates de novo-development of antimicrobial peptides. Nat Commun 2023;14:7197.
[20] Lv X et al. The C-terminal sequences of porcine thrombin are active as antimicrobial peptides. Chem Biol Drug Des 2016;88:905–14.
[21] Lv X et al. The C-terminal sequences of porcine thrombin are active as antimicrobial peptides. Chem Biol Drug Des 2016;88:905–14.
[22] Clark S, Jowitt TA, Harris LK, Knight CG, Dobson CB. The lexicon of antimicrobial peptides: a complete set of arginine and tryptophan sequences. Commun Biol 2021;4:605.
[23] Saint Jean KD et al. Effects of hydrophobic amino acid substitutions on antimicrobial peptide behavior. Probiotics Antimicrob Proteins 2018;10:408–19.
[24] Yao A et al. Progressive design of a ranatuerin-2 peptide from amolops wuyiensis: enhancement of bioactivity and in vivo efficacy. Antibiotics 2023;13:5.
[25] Medina-Franco JL et al. Characterization of activity landscapes using 2D and 3D similarity methods: consensus activity cliffs. J Chem Inf Model 2009;49:477–91.
[26] Peltason L, Bajorath J. SAR index: quantifying the nature of structure–activity relationships. J Med Chem 2007;50:5571–8.
[27] Guha R. Exploring uncharted territories: predicting activity cliffs in structure-activity landscapes. J Chem Inf Model 2012;52:2181–91.
[28] Heikamp K, Hu X, Yan A, Bajorath J. Prediction of activity cliffs using support vector machines. J Chem Inf Model 2012;52:2354–65.
[29] Tancik M et al. Fourier features let networks learn high frequency functions in low dimensional domains. Adv Neural Inf Proces Syst 2020;33:7537–47.
[30] Jacot A, Gabriel F, Hongler C. Neural tangent kernel: Convergence and generalization in neural networks. Adv Neural Inf Proces Syst 2018;31.
[31] Xia J, et al. in Thirty-seventh Conference on Neural Information Processing Systems (2023).
[32] Rahaman N. et al. in International Conference on Machine Learning 5301-5310 (PMLR, 2019).
[33] Ronen B, Jacobs D, Kasten Y, Kritchman S. The convergence rate of neural networks for learned functions of different frequencies. Adv Neural Inf Proces Syst 2019;32.
[34] Valdar WS. Scoring residue conservation. Proteins Struct Funct Bioinf 2002;48:227–41.
[35] Wang G, Li X, Wang Z. APD3: the antimicrobial peptide database as a tool for research and education. Nucleic Acids Res 2016;44:D1087–93.
[36] Shi G, et al. DRAMP 3.0: an enhanced comprehensive data repository of antimicrobial peptides. Nucleic Acids Res **50**, D488-d496 (2022).
[37] Pirtskhalava M et al. DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics. Nucleic Acids Res 2021;49:D288–97.
[38] Piotto SP, Sessa L, Concilio S, Iannelli P. YADAMP: yet another database of antimicrobial peptides. Int J Antimicrob Agents 2012;39:346–51.
[39] Witten J, Witten Z. Deep learning regression model for antimicrobial peptide design. BioRxiv 2019;692681.
[40] Pendleton JN, Gorman SP, Gilmore BF. Clinical relevance of the ESKAPE pathogens. Expert Rev Anti Infect Ther 2013;11:297–308.
[41] Lin Z et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science 2023;379:1123–30.
[42] Nijkamp E, Ruffolo JA, Weinstein EN, Naik N, Madani A. Progen2: exploring the boundaries of protein language models. Cell Syst 2023;14. 968–978. e963.
[43] Schissel CK et al. Deep learning to design nuclear-targeting abiotic miniproteins. Nat Chem 2021;13:992–1000.
[44] Deng J et al. A systematic study of key elements underlying molecular property prediction. Nat Commun 2023;14:6395.
[45] Fernandez-Diaz, R. et al. AutoPeptideML: Automated machine learning for building trustworthy peptide bioactivity predictors. bioRxiv, 2023.2011. 2013.566825 (2023).
[46] Altschul SF et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–402.
[47] Fernandez-Lopez S et al. Antibacterial agents based on the cyclic D, L-α-peptide architecture. Nature 2001;412:452–5.
[48] Koo YS et al. Structure–activity relations of parasin I, a histone H2A-derived antimicrobial peptide. Peptides 2008;29:1102–8.
[49] Mouton JW et al. MIC-based dose adjustment: facts and fables. J Antimicrob Chemother 2018;73:564–8.
[50] (2024).
[51] Hauser M, Steinegger M, Soding J. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. Bioinformatics 2016;32:1323–30.
[52] Baylon JL et al. PepSeA: peptide sequence alignment and visualization tools to enable lead optimization. J Chem Inf Model 2022;62:1259–67.
[53] Hayes T, et al. Simulating 500 million years of evolution with a language model. bioRxiv, 2024.2007. 2001.600583 (2024).
[54] Kaplan J, et al. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361; 2020.
[55] Tian X, Sun F, Zhou XR, Luo SZ, Chen L. Role of peptide self-assembly in antimicrobial peptides. J Pept Sci 2015;21:530–9.
[56] Wu Z et al. MoleculeNet: a benchmark for molecular machine learning. Chem Sci 2018;9:513–30.
[57] Yang K et al. Analyzing learned molecular representations for property prediction. J Chem Inf Model 2019;59:3370–88.