



OPEN AI-driven antimicrobial peptide characterization unveils novel motifs for drug design

Sarala Padi¹✉, Kinjal Mondal^{1,3,5}, David P. Hoogerheide², Frank Heinrich^{2,4}, Mihaela Mihailescu³, Jeffery B. Klauda^{5,6} & Antonio Cardone¹

Antibiotics have been developed to effectively target and eliminate bacteria, but the rise in antimicrobial resistance (AR) complicates the treatment of certain infections. To address this issue, researchers have explored antimicrobial peptides (AMPs) that disrupt bacterial membranes. A promising method for this exploration is motif-based analysis, which identifies hidden patterns in AMPs to better understand their mechanism of action. While existing methods rely on expert knowledge, incorporating topic models can enhance analysis by revealing the contextual relationships between sequence elements. This is complemented by a data analytics tool designed to analyze AMP motifs and their biochemical properties. Such integration allows for the extraction of valuable motifs and the development of a robust data analytics module for predicting membrane activity. Additionally, we evaluated the biological relevance of motifs by extracting biochemical features, making structural predictions via Evolutionary Scale Modeling (ESM). Our results indicate that topic model-derived motifs are strongly associated with antimicrobial activity and demonstrate lower minimum inhibitory concentration values and capture contextual information more effectively than traditional frequency-based motifs. We also performed a comparative analysis between the two approaches regarding motif evolution, sequence-level attributes, and entropy measures, ultimately contributing to ongoing efforts to combat AR.

Keywords Antimicrobial resistance, Antimicrobial peptides, Drug design, Motif extraction, Topic model, Minimum inhibitory concentration

Antibiotics have been developed to treat bacterial infections, significantly improving health and life expectancy. However, some bacterial strains develop resistance to antibiotics, known as antibacterial resistance (AR)¹. AR increases the risk of severe illness and poses a global public health threat, causing millions of deaths annually^{2–4}. In response, the World Health Assembly adopted a global action plan to combat AR by investing in new medicines and interventions⁵.

To effectively combat infections caused by AR, scientists are focusing on antimicrobial peptides (AMPs) as promising candidates for the next generation of antibiotics^{6,7}. AMPs, naturally found in all living organisms, play a crucial role in defending against fungi, viruses, and bacteria^{8–11}. These peptides, consisting of 10 to 60 amino acids, can be engineered to disrupt bacterial cell membranes⁹. The effectiveness of AMPs relies on their unique biochemical properties, particularly the arrangement of structural features like α -helices and β -sheets. To advance AMP research, a strong interdisciplinary approach that integrates biology, material science, chemistry, and bioinformatics is essential. Such collaboration is vital for developing innovative solutions to combat AR^{6,8,9,12}.

Motifs are short subsequences linked to secondary structures, like helices. Identifying new sequences often shows limited similarities to known ones, emphasizing the importance of motif analysis for understanding their roles¹³. By characterizing motifs, we gain insights into peptide functionality, particularly when structurally distinct peptides serve similar functions. They play crucial roles in biological processes such as gene expression regulation, protein design, and functional annotation^{14–16}. Although regular expressions are often used for

¹Information Technology Laboratory (ITL), NIST, Gaithersburg, MD 20899, USA. ²NIST Center for Neutron Research (NCNR), NIST, Gaithersburg, MD 20899, USA. ³Institute for Bioscience and Biotechnology Research (IBBR), UMD, Rockville, MD 20850, USA. ⁴Department of Physics, Carnegie Mellon University, Pittsburgh, PA 15213, USA. ⁵Institute for Physical Science and Technology, Biophysics Program, UMD, College Park, MD 20742, USA. ⁶Department of Chemical and Biomolecular Engineering, UMD, College Park, MD 20742, USA. ✉email: sarala.padi@nist.gov

motif identification, they require prior knowledge and may miss important motifs. To improve motif discovery, complementary approaches like statistical models can be helpful. This study uses topic models to automatically extract motifs from protein sequences, enhancing our understanding of protein functionality.

Recent advances in computational biology have led to notable progress in the development of AI models, including large language models, such as ESM¹⁷, ProGen¹⁸, and ProtBERT¹⁹, as well as structure prediction models, including AlphaFold²⁰ and ESMfold¹⁷. Although these models excel in function, sequence generation, and structure prediction, their reliance on extensive datasets for fine-tuning can limit adaptability and the extraction of meaningful insights. Additionally, these methods can be biased, particularly in recognizing novel motifs that are less represented in the data. Inspired by techniques from natural language processing, topic models effectively reveal hidden structures and relationships without labeled examples, offering better interpretability than traditional protein-based models²¹. Previous studies have shown that topic models provide deeper insight into protein-protein interactions, identify meaningful patterns, and facilitate a deeper understanding of biological systems^{22–24}.

Several computational approaches have been proposed for motif discovery, ranging from traditional sequence-based algorithms to AI-driven frameworks, but each has limitations when it comes to AMPs. For example, AMPlify²⁵ highlights key residues using attention mechanisms but does not define functional motifs. Similarly, MotifQuest²⁶ and Gemoda²⁷, while useful for mass spectrometry data and k-mer clustering, lack the contextual understanding needed for AMP analysis.

In DNA motif discovery, algorithms like MEME, Weeder, and AlignACE have been validated on synthetic datasets with known motifs, enabling clear benchmarking²⁸. However, the absence of comparable ground truth motifs for AMPs, which derive their functional motifs from sequence activity, presents an opportunity for developing tailored benchmarking approaches. The BML web server²⁹ and AI frameworks like PLPTP³⁰ offer insights into motif discovery, yet they tend to focus on regulatory DNA sequences and toxicity prediction, limiting their application to AMPs. This highlights a key challenge: the lack of established ground truths for biologically significant motifs complicates the assessment of their functional relevance. Similarly, machine learning-based AMP generation pipelines often rely on extracting recurring motifs from sequences predicted as antimicrobial. These motifs are then ranked by frequency or statistical association (e.g., lasso regression) and used as building blocks for new peptide design³¹. While this provides interpretability, the approach treats motifs as isolated fragments and ignores the broader sequence context.

To address these gaps, we propose a topic modeling framework that treats motifs as latent themes within AMP sequences, providing a more context-aware representation. This approach allows for the discovery of coherent patterns critical for antimicrobial activity. Our AI-driven framework designed to extract meaningful motifs from AMPs by dividing sequences into smaller subsequences or k-mers. This thematic approach aids in revealing critical patterns essential for antimicrobial activity.

Furthermore, we develop a data analysis framework to highlight key motifs. These motifs represent promising candidates for AMP design and exhibit functional diversity across various thematic clusters. Our analysis identifies motifs and computes properties to improve biological understanding by mapping sequence level properties to motifs and computing motif level properties.

We develop a data analysis framework that highlights key motifs, which are promising candidates for AMP design. These motifs show functional diversity across different thematic clusters and enhance biological understanding by linking sequence properties to motifs and assessing their characteristics. We investigate how motif sizes and different topic counts affect motif extraction and data clarity. Our findings indicate that motifs derived from topic models are more diverse and capture contextual information more effectively than those based on frequency from databases. Additionally, we compared both types of motifs in terms of their relevance to motif evolution, sequence attributes, and entropy measures. This comprehensive approach enhances the understanding of antimicrobial research and contributes positively to tackling antimicrobial resistance.

Methods

The identification of hidden patterns within antimicrobial peptide sequences offers significant opportunities for understanding their structural functionalities and mechanisms of action, particularly in terms of membrane binding, insertion, and disruption. These patterns or subsequences represent critical structural elements that are essential for the biological activity of AMPs. However, identifying such subsequences is inherently complex because of the vast and diverse range of AMPs available in current biological databases. In many cases the identification process relies heavily on the expertise of domain specialists who manually examine the sequences, drawing on their knowledge of specific AMP classes and associated biophysical properties.

We propose a framework for motif analysis in AMPs, focusing on k-mers—short subsequences crucial for antimicrobial activity. Our framework includes three modules: (1) encoding or representing AMPs, (2) using topic models to generate distributions, and (3) mapping extracted motifs to relevant properties. This structured approach aims to uncover the biological significance of identified motifs. We adapt Latent Dirichlet Allocation (LDA), a proven model in natural language processing, for motif discovery in biological sequences. By developing a pipeline that applies LDA to AMP sequences and validating the motifs, we can reveal biologically relevant patterns and enhance our understanding of membrane activity. This work extends the application of LDA and opens new research avenues in the field of antimicrobial peptides.

Encoding of AMPs: k-mer generation

Sequence encoding is a crucial step in developing topic models. In this analysis, we used a k-mer representation to encode AMPs, similar to n-grams in natural language processing. A k-mer is a contiguous subsequence of length k from an amino acid sequence³². To compute k-mers, we chose two main parameters: the extraction method and the length of the k-mers. In our study, we extracted k-mers using overlapped window method to

capture detailed information. Shorter k-mers help identify local motifs, while longer ones provide a broader view but can lead to data sparsity. Finding the optimal “k” is important for capturing the biological relevance of AMPs. Thus, we analyzed k-mer lengths from 2 to 20. This range includes short amino acid patterns (2–6 residues) significant for charge or hydrophobicity in antimicrobial activity and longer motifs (10–20 residues) corresponding to structural elements of AMPs. We also developed metrics to assess k-mer length significance in motif analysis. Figure 1 shows an illustration of k-mer generation for a given sequence with length of 12 and a k-mer length of 3.

Motif extraction: LDA topic model vs frequency-based approaches

LDA is a powerful probabilistic model used to uncover hidden topics within discrete data, such as biological sequences represented as k-mers. As an unsupervised model, it effectively groups co-occurring k-mers that exhibit functional similarities into distinct topics. The topic modeling is a powerful concept, embedding documents as a mixture of topics and defining topics as a mixture of words. In our adaptation of this analysis for motif extraction, we treat sequences as “documents” and k-mers as “words”. Each sequence is viewed as a mixture of latent topics, defined by unique probability distributions³³. LDA groups k-mers into topics based on a certain probability, and we select the top ten motifs that exhibit high coherence within each topic.

In our study, we used Gensim tool³⁴ to build the LDA model, applying techniques like variational inference. Training of LDA model, outlined in Algorithm 1, includes bag-of-k-mers similar to the bag-of-words representation in NLP. The initial random assignments of k-mers to topics are refined through iterative reassignment based on topic prevalence and k-mer frequency until convergence is achieved. As shown in Fig. 2, after training, the LDA topic model assigns a probability score to each motif within a topic, indicating how representative that motif is of the topic. By utilizing the co-occurrence of k-mers within the bag-of-k-mers, LDA successfully captures contextual relationships and uncovers hidden patterns in the data. In this study, we selected the top 10 motifs from each topic based on their probability scores to balance interpretability and coverage. The number of motifs considered is a design choice that can be adapted depending on the application. For example, when designing AMPs, a larger number (e.g., top 20 motifs) may be used to enhance sequence diversity, while avoiding motifs with redundant or undesired amino acid patterns such as multiple tryptophan (W) residues.

In topic modeling, the selection of optimal number of topics is crucial, as too few can oversimplify the model, while too many can complicate the analysis. In addition, the number of topics does not directly indicate biological relevance due to the lack of ground truth data. With the true number of motifs unknown, we explored topic models within the 2 to 30 topics range, using the lower bound for minimal separation and the upper bound based on dataset size and motif diversity. We selected the optimal number of topics by maximizing coherence score and conducted additional biological validation at the motif-property level. The selection of LDA hyperparameters (α and η) is vital for model stability and interpretability while ensuring the biological validity of extracted motifs. In this study, the values of α and η are fixed at 0.01, using a fixed seed.

By creating a bag of k-mers, we compute frequency-based motifs. This non-probabilistic approach suggests that the most frequent subsequences represent the motifs effectively. Algorithm 2 illustrates the frequency-based motif extraction process for the AMP analysis. This method allows us to compare motifs generated through LDA models, showing how LDA captures contextual information compared to frequency-based motifs. Such comparisons yield insights into the topic model’s effectiveness in uncovering hidden patterns in the data.

The LDA topic model is an unsupervised learning approach that analyzes the relationships between amino acids in sequences. To understand the relevance of these motifs, we focused on three key physicochemical properties of AMPs: (i) isoelectric point, indicating charge; (ii) hydrophobic character; and (iii) secondary structure. Using the Bio-Python tool³⁵, we calculated these properties, which enhanced our understanding of their biological significance. A summary of these properties is available in Section 1 of the Supplementary Material (SI).

Metrics

We systematically evaluated our AI-driven motif extraction pipeline by leveraging key metrics. Firstly, we utilized the “Coherence Score” to measure the average cosine similarity of motif pairs within each topic, which enabled us to effectively assess and select the most suitable topic model. We also measured entropy to evaluate motif variability, helping us identify the topics of interest and analyze topics by analyzing Minimum Inhibitory Concentration (MIC) values. To enrich our findings, we considered several important properties, including the normalized hydrophobic moment (HM), isoelectric point (IP), GRAVY score, and secondary structure.

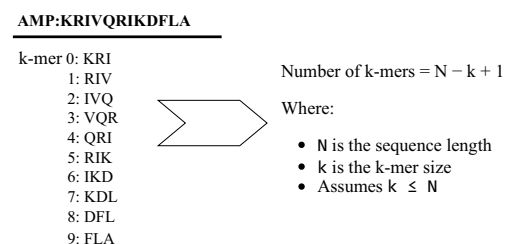


Fig. 1. An illustration of k-mer generation for a given sequence with a length of 12 and a k-mer length of 3. The default overlap length is 1.

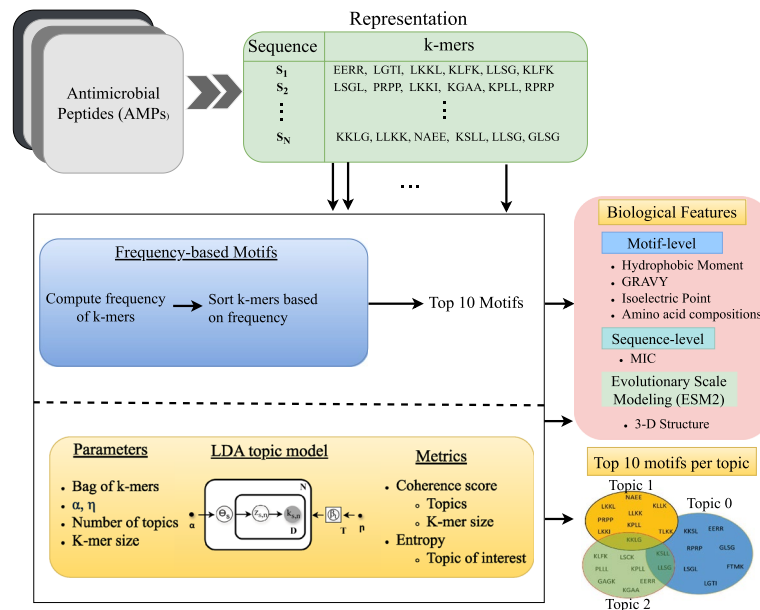


Fig. 2. AI-driven motif extraction framework for drug design analysis. This framework extracts k-mers and uses a bag-of-k-mers to represent AMPs. We compute the frequency of k-mers to identify the top 10 motifs from a sequence database. Using these bag-of-k-mers, we train a LDA model, which views each sequence as a mixture of topics, with each topic comprising various motifs, resulting in distinct motif clusters. Key LDA parameters, such as the number of topics and k-mer size, are chosen based on coherence scores, while the topic of interest is identified using an entropy measure. Parameters such as α and η regulate the diversity and coherence of motifs in each topic and are set to 0.01 for motif analysis. The model generates clusters of motifs. These clusters are analyzed at both the motif and sequence levels by calculating hydrophobic moment, GRAVY, isoelectric point, amino acid composition, and MIC at the sequence level. Additionally, we utilize evolutionary scale modeling (ESM2) to compute the 3-D structures of these motifs, enhancing our ability to evaluate their biological significance. The analysis also includes motifs extracted through frequency-based method.

Additionally, we analyzed the frequency of amino acid residues, quantifying positively charged amino acids that enhance electrostatic interactions with bacterial membranes, as well as hydrophobic residues (L, I, F, V, W) and polar amino acids (S, T, Y) that affect solubility. Finally, we examined the roles of glycine (G) and proline (P) in promoting flexibility in antimicrobial peptide structures³⁶. Refer to Sections 2 and 3 in the SI for more information.

- 1: **Input:** Set of sequences $S = \{s_1, s_2, \dots, s_N\}$, k-mer length k , number of topics T , number of top motifs per topic T_{top}
- 2: **Output:** Set of T_{top} motifs for each topic $t \in \{1, \dots, T\}$

▷ **Bag-of-k-mer representation**

- 3: Initialize corpus $\mathcal{D} \leftarrow \emptyset$
- 4: **for** $i = 1$ to N **do**
- 5: Initialize sequence $d_i \leftarrow []$
- 6: **for** $j = 0$ to $|s_i| - k$ **do**
- 7: Extract $m = s_i[j : j + k]$
- 8: Append m to d_i
- 9: **end for**
- 10: Add d_i to corpus \mathcal{D}
- 11: **end for**

▷ **LDA training and topic extraction**

- 12: Train LDA model on corpus \mathcal{D} to obtain:
- 13: Topic-motif distributions $\phi_t(m) = P(m | z = t)$ for each topic $t \Rightarrow$ probability of 'm' given that the generating topic $z = t$, where z is a latent topic variable
- 14: **for** $t = 1$ to T **do**
- 15: Sort motifs m in descending order of $\phi_t(m)$
- 16: Select T_{top} motifs from topic t
- 17: **end for**
- 18: **return** T_{top} motifs from each topic t

Algorithm 1. LDA-based motif extraction.

```

1: Input: Set of topics  $T = \{t_1, t_2, \dots, t_n\}$ , each with a set of motifs  $M_i = \{m_{i1}, m_{i2}, \dots, m_{i10}\}$ 
2: Output: Mean property value per topic
3: for each topic  $t_i \in T$  do
4:   for each motif  $m_{ij} \in M_i$  do, where  $1 \leq j \leq 10$ 
5:     Compute IP, GRAVY, HM, HTF
6:   end for
7:   Compute the average properties over the top 10 motifs within topic  $t_i$ 
8: end for

```

Algorithm 2. Frequency-based motif extraction.

Database

To develop topic models that uncover motifs and assess the influence of significant motifs on the antimicrobial activity of peptides, we created a comprehensive dataset focused on linear AMPs composed of standard amino acids. Non-standard amino acids (B, J, O, U, X, and Z) were excluded. We compiled a dataset that includes MIC values standardized to $\mu\text{mol/L}$. For the ranges of MIC values, we calculated the mean, and for values greater than or less than a specific threshold, we used that threshold. When multiple MIC measurements were available for an identical sequence, the mean value is used to account for experimental variability otherwise each sequence is treated as unique. Furthermore, we ensured that any molecular weight unit was accurately converted to $\mu\text{mol/L}$ by utilizing the standard molecular masses of the peptides.

Table S1 lists the databases and online sources used for the data collection. The data were initially sourced from three different databases: (i) GRAMPA³⁷, which includes information from other databases such as the Antimicrobial Peptide Database (APD)³⁸, Database of Antimicrobial Activity and Structure of Peptides (DBAASP)³⁹, YADAMP⁴⁰, and DRAMP⁴¹; (ii) StarPep^{42,43}; and (iii) DBAASP3³⁹. In this study, we curated a comprehensive dataset of 5,860 sequences with MIC values against the *E.Coli* target. Figure 3 shows the length distribution of the AMP sequences used for motif analysis.

Results

Optimal k-mer size and number of topics

Figure 4 illustrates the average coherence score for k-mer lengths varying from 2 to 20, averaged over topic counts ranging from 2 to 30. In this analysis, we averaged the coherence scores over the topics corresponding to each k-mer length. As illustrated in Fig. 4, we identified peaks at the k-mer lengths of 4, 14, and 18. This suggests that the LDA topic model performs consistently at these specific k-mer sizes, regardless of the number of topics. Notably, the model reliably produced results for motifs likely associated with helical structures, as helices are typically defined by 4 amino acid residues. This is particularly relevant, given that many AMPs are characterized by their helical and amphipathic properties. Moreover, the peak at k-mer lengths 14, and 18 indicates strong reliability for motifs associated with beta sheets.

After selecting k-mer sizes of 4, 14, and 18, we use the coherence score to determine the optimal number of topics for each length. As shown in Fig. S10 in the Supplementary Information, the optimal topics for k-mer lengths of 4, 14, and 18 are 2, 2, and 4, respectively. We further analyze these topics to identify motifs and assess their biological significance in AMP analysis.

Motif overlap: LDA-derived vs frequency-based motifs

Table 1 shows motifs derived from LDA and frequency-based methods for 4-mers, 14-mers, and 18-mers. To quantify the motifs extracted using LDA and frequency-based methods, we analyzed the overlap of 4-, 14-, and 18-mer motifs both within LDA and frequency-based methods, as well as across the topics generated by these methods.

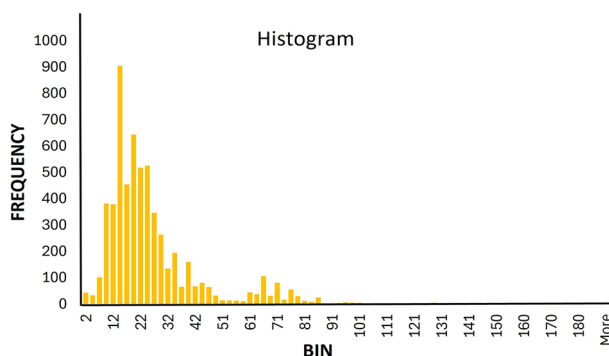


Fig. 3. Shows the length distribution of the AMP sequences. As shown, the majority of sequences have lengths between 12 and 32 residues.

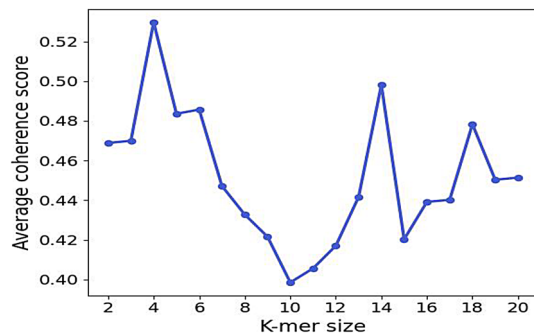


Fig. 4. Shows the average coherence score for k-mer lengths from 2 to 20, with topic counts ranging from 2 to 30. We averaged the scores based on 2 to 20 topics for each k-mer length to select the optimal k-mer sizes for motif analysis.

Method	k-mer size	Topic	Top 10 motifs
LDA-derived	4	0	NAEE, LKKL, KLLK, TLKK, KKLL, PRPP, LLKK, KPLL, KKLK, LKKI
		1	LGTI, KKSLL, KSLK, EERR, TLKK, LSLG, FTMK, RPRP, LLSG, GLSG
	14	0	LLSGILGAGKHIVC, LSGILGAGKHIVCG, SGILGAGKHIVCGL
			GLLSGILGAGKHIV, GILGAGKHIVCGLS, ILGAGKHIVCGLSG
			LGAGKHIVCGLSGL, GAGKHIVCGLSGLC, LLIGAGKSAQSVL
		ALNAAKSAGVSVLN	
		1	TGLELMACKITNQC, KTGLELMACKITNQ, GKTGLELMACKITN
			FTLIKGAAKLIGKT, TLIKGAAKLIGKTV, GLFTLIKGAAKLIG
	LFTLIKGAAKLIGK, GKTVAKEAGKTGLE, VAKEAGKTGLELMA		
	KTVAKEAGKTGLEL		
	18	0	LLSGVLGVGKKIVCGLSG, SGVLGVGKKIVCGLSGLC, GLLSGVLGVGKKIVCGLS
			LSGVLVGKKIVCGLSGL, ALNAAKSAGVSVLNSLSC, LNAAKSAGVSVLNSLCK
			SAGVSVLNSLCKLSKTC, AAKSAGVSVLNSLCKLS, AKSAGVSVLNSLCKLSK
			KSAGVSVLNSLCKLSKT
		1	KTVAKEAGKTGLELMACK, IGKTVAKEAGKTGLELMA, GKTVAKEAGKTGLELMAC
			LIGKTVAKEAGKTGLELM, KEAGKTGLELMACKITNQ, AKEAGKTGLELMACKITN
			TVAKEAGKTGLELMACKI, KGAAKLIGKTVAKEAGKT, FTLIKGAAKLIGKTVAKE
			GAAKLIGKTVAKEAGKTG
2		LLSGILGAGKHIVCGLSG, GLLSGILGAGKHIVCGLS, LSGILGAGKHIVCGLSGL	
		SGILGAGKHIVCGLSGLC, LIGAGKSAQSVLKTLS, SAAQSVLKTLSCKLSNDC	
		KSAQSVLKTLSCKLSND, IGAGKSAQSVLKTLSCK, GAGKSAQSVLKTLSCKL	
		LLIGAGKSAQSVLKTLS	
3	LKGCWTKSIPPKPCFGK, ALKGCWTKSIPPKPCFGK, AALKGCWTKSIPPKPCFG		
	CVYAYVRVGVLVRYRRC, IGKEVGMVIRTGIDVAG, VGMVIRTGIDVAGCKIK		
	EVGMVIRTGIDVAGCKI, KEVGMVIRTGIDVAGCK, MDVIRTGIDVAGCKIKGE		
	GKEVGMVIRTGIDVAGC		
Freq-based	4	-	FFLG, LFFL, LSLC, LLLL, LLLF, EERR, FLGT, LFFF, LGTI, KSLK
		-	MFTLKKSLLLLFLL, FTLKKSLLLLFLLG, LLLFLLGTINLSL
	14	-	LKKSLLLLFLLGTI, TLKKSLLLLFLLGT, LLLFLLGTINLSL
		-	KKSLLLLFLLGTIN, KSLLLLLFLLGTINL, SLLLLFLLGTINLS
	18	-	EERRDEEVAKMEE
		-	KKSLLLLFLLGTINLSL, ETNAEEERRDEEVAKMEE, DETNAEEERRDEEVAKME
		-	TNAEEERRDEEVAKMEEI, EEERRDEEVAKMEEIKRG, MFTLKKSLLLLFLLGTIN
		-	CQDETNAEEERRDEEVAK, QDETNAEEERRDEEVAKM, AEEERRDEEVAKMEEIKR
NAEEERRDEEVAKMEEIK			

Table 1. Shows frequency-based and LDA-derived 4-, 14-, and 18-mer motifs. The optimal number of topics was determined based on the LDA model’s highest coherence score. Our analysis focuses only on the top 10 motifs identified through LDA and frequency-based methods. Note: “-” means unlike LDA-based model, there are no topic assignments for the frequency-based analysis.

Method	K-mer size	Contained In	Motifs included (s/m)	Motifs containing them (r/k)
Frequency-based	4	14	10/10	10/10
	4	18	10/10	10/10
	14	18	10/10	7/10
LDA model	4	14	3/20	5/20
	4	18	3/20	8/20
	14	18	17/20	14/40

Table 2. Overlap of motifs within LDA-based and frequency-based motifs of different lengths. Abbreviation: #Motifs Included: Number of shorter length motifs are contained in motifs of longer length, #Motifs containing Them: Number of Longer length motifs actually contain shorter length motifs. s/m: 's' number of shorter length motifs out of 'm' are contained in longer length motifs. r/k: 'r' number of longer length motifs out of 'k' actually containing shorter length motifs 's'. For example, s/m:3/20 indicates that 3 4-mers out of 20 are contained in any of 18-mers; r/k: 5/20: indicates that those 3 of shorter length k-mers appear in 5 out of 20 14-mer motifs.

Our analysis in Table 2 shows that frequency-based motifs exhibit impressive overlap: all 4-mer motifs are contained within both the 14-mer and 18-mer sets (10/10), and a significant number of 14-mers (7/10) are also found in the 18-mer set, as shown in Figs. S1, S2, and S3 in the SI. In contrast, LDA-derived motifs show less overlap, with only three out of 20 4-mers appearing in the 14-mer and the 18-mer sets. This suggests that LDA captures unique contextual information, as the overlapping motifs are distributed among only five of the 14-mers and eight of the 18-mers. Additionally, while 17 of the 20 LDA 14-mers are present in the 18-mer set, only 14 of the 40 include them (see Figs. S4, S5, and S6). Furthermore, minimal redundancy between LDA-derived and frequency-based motifs enhances their complementary nature, with only two of the LDA 4-mers matching the frequency-based list and no overlap in the 14-mer or 18-mer sets. Overall, these findings highlight that while frequency-based motifs identify prevalent sequences, LDA excels in uncovering unique, context-specific motifs. By leveraging both techniques, we can gain a more comprehensive understanding of the motif structure and sequence functionality in biological datasets, ultimately enriching our insights into biological processes.

Entropy-guided topic selection for motif discovery

In this study, we examined the use of entropy measures to achieve two main goals: 1) identifying topics of interest and 2) distinguishing between diverse and conserved motifs. While the coherence score is useful for assessing topic quality, it has limitations for direct comparisons. Therefore, we implemented entropy measures to clarify motifs from the topic model and highlight areas for further research. Lower entropy values indicate highly conserved motifs that are likely stable and significant evolutionarily, though not all conserved motifs have distinct biological functions. In contrast, higher entropy values suggest increased variability or noise, while motifs with intermediate values often relate to known AMP domains. In Table S2 of the SI, we present detailed entropy values for motifs from frequency-based methods alongside those from the LDA model. Frequency-based motifs exhibited lower entropy (below 2.5) for both 4-mer and 14-mer motifs, indicating a tendency towards repetitive patterns that may limit functional diversity. Conversely, the LDA model revealed varying mean entropy values across 4-mer, 14-mer, and 18-mer motifs, highlighting significant diversity and complexity in the identified motifs. Notably, the entropy of 18-mer motifs from the LDA model in topic 3 was higher than that of frequency-based motifs, indicating a richer exploration of sequence variability and potential functional diversity. Although this trend was less consistent for 4-mer motifs in topic 0, it suggests promising areas for future research.

Motif analysis

As mentioned in Section , we identified optimal k-mer sizes of 4, 14, and 18 based on LDA model coherence scores, resulting in 2 topics for k-mer sizes 4 and 14, and 4 topics for k-mer size 18. To explore the biological significance of the motifs related to antimicrobial, we used two approaches: 1) calculating motif-level properties and 2) mapping sequence-level attributes to motifs.

- **Motif-level property comparison:** In this approach, we assess motif properties across topics to uncover their biological significance. The algorithm for quantifying motif relevance, based on HM, GRAVY, IP, Helix and Turn Fraction (HTF), is outlined in Algorithm 3.
- **MIC sequence-level comparison:** Sequence-level properties were mapped using motif-level aggregation and topic-level summarization. This approach helps in understanding how different motifs influence MIC values across topics. The procedure for quantifying motif relevance per topic through MIC association is described in Algorithm 4.

```

1: Input: Set of sequences  $S = \{s_1, s_2, \dots, s_N\}$ , k-mer length  $k$ , number of top motifs  $T$ 
2: Output: Set of Top  $T$  frequent motifs
3: Initialize an empty frequency dictionary  $f(m) \leftarrow 0$ 

4: for  $i = 1$  to  $N$  do
5:   Let  $s_i$  be the  $i$ -th sequence
6:   for  $j = 0$  to  $|s_i| - k$  do
7:     Extract  $m = s_i[j : j + k]$ 
8:      $f(m) \leftarrow f(m) + 1$ 
9:   end for
10: end for
11: Sort motifs  $m$  in descending order of frequency  $f(m)$ 
12: return Top  $T$  motifs with highest  $f(m)$ 

```

▷ k-mer extraction
▷ frequency computation

Algorithm 3. Motif-level property comparison: quantifying motif relevance per topic using HM, GRAVY, IP, HTF properties computed at motif level.

```

1: Input: Set of topics  $T = \{t_1, t_2, \dots, t_n\}$ , each with a set of ten motifs  $M_i = \{m_{i1}, m_{i2}, \dots, m_{i10}\}$ 
2: Output: Mean MIC value per topic
3: for each topic  $t_i \in T$  do
4:   for each motif  $m_{ij} \in M_i$  do, where  $1 \leq j \leq 10$ 
5:     Find all sequences that contain motif  $m_{ij}$  (count each sequence only once)
6:     Retrieve MIC values for the matched sequences
7:     Compute the mean MIC for motif  $m_{ij}$ 
8:   end for
9:   Compute the average MIC over the top 10 motifs within topic  $t_i$ 
10: end for

```

Algorithm 4. MIC sequence-level comparison: quantifying motif relevance per topic by MIC association at sequence level.

Comparison 1: motif-level properties

We calculated three different properties at the motif level: isoelectric point, hydrophobic moment, and GRAVY. These properties were used to analyze the motifs extracted using the LDA-based and frequency-based methods. Additionally, we generated 3D structures for longer motifs (14-mer and 18-mer) using ESM Fold software. To enhance our understanding of the amino acid compositions of the motifs derived from both methods, we determined the frequency of the top motifs and visually assessed the differences and significance of the motifs extracted using each approach.

4-mer motifs: From Fig. 5, it is evident that the motifs in Topic 0 display a higher IP than those in Topic 1, which have a broader range typically spanning from 5 to 12. Conversely, motifs based on frequency analysis exhibited lower IP values than those derived from topics 0 and 1. It is interesting to note that motifs generated

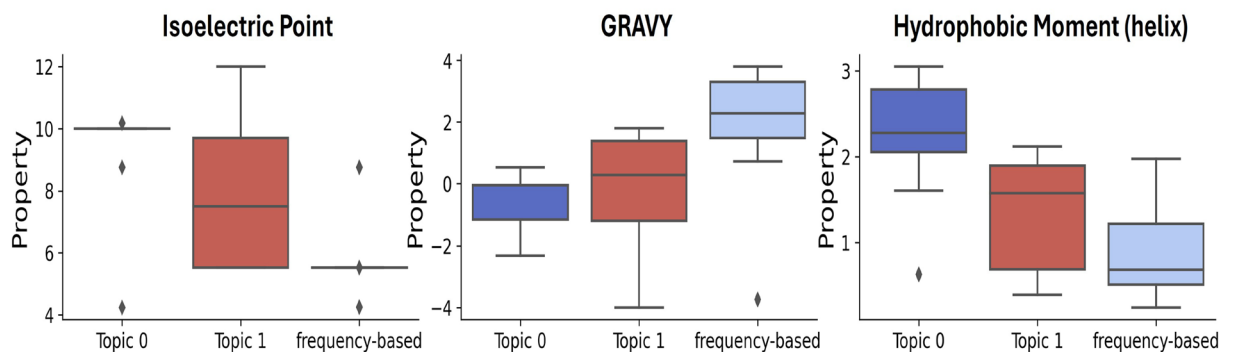


Fig. 5. The distribution of motif properties for the top 10 4-mer: LDA model-derived motifs and frequency-based motifs. Motifs in topic 0 show a higher IP than those in topic 1, indicating that topic 0 motifs are more basic, while frequency-based motifs tend to be more acidic. In terms of hydrophobic moment, topic 0 motifs show higher values as compared to topic 1 motifs, whereas frequency-based motifs display lower values than both topics. This suggests that topic modeling-derived motifs are more amphipathic. Additionally, the lower GRAVY index of both topic 0 and topic 1 motifs compared to frequency-based ones opens new avenues for exploring their interactions in biological contexts.

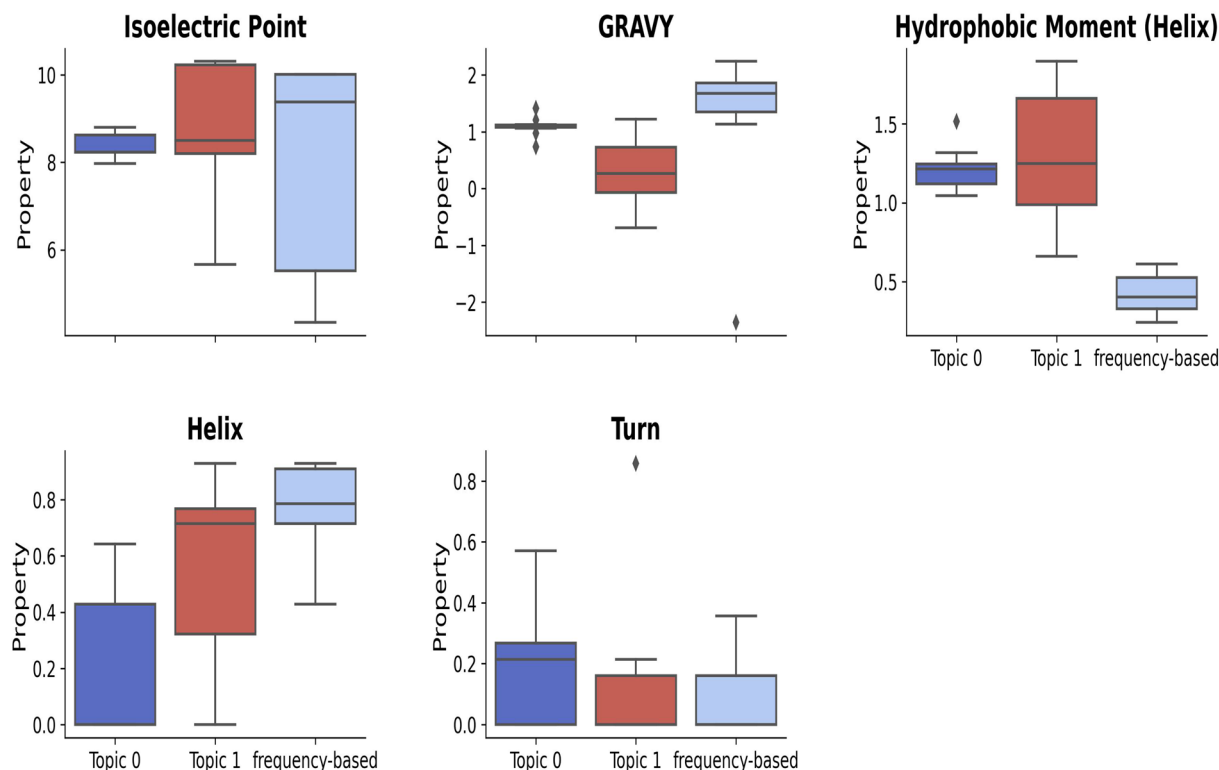


Fig. 6. An overview of motif-level feature distributions for the top 10 14-mer LDA- and frequency-based motifs related to *E. Coli*. The motifs in topics 0 and 1 have IP values above 7, indicating they are predominantly basic, while frequency-based motifs show a broader spectrum, including both basic and acidic. Frequency-based motifs also have a higher GRAVY index, indicating greater hydrophobicity, with topic 0 motifs showing a particularly high index. Both topics 0 and 1 exhibit higher hydrophobic moments, but topic 0 has less variability. In contrast, frequency-based motifs have a lower hydrophobic moment. Regarding helix and turn fractions, topics 0 and 1 present unique helix fraction ranges, while frequency-based motifs tend to have higher helix fractions. Both motif types exhibit similar trends in turn fractions, highlighting the complementary insights provided by topic modeling and frequency analysis in understanding motif characteristics.

through LDA are basic, whereas frequency-based motifs are on the more neutral side. Additionally, the motifs in Topic 0 had a greater HM than those in Topic 1. In contrast, frequency-based motifs demonstrated a lower HM than both Topics 0 and 1, indicating that motifs derived from the topic modeling approach are more amphipathic than frequency-based motifs. Furthermore, both Topic 0 and Topic 1 motifs exhibited a lower GRAVY index than frequency-based motifs.

14-mer motifs: Figure 6 shows the motif-level feature distributions for the top 10 14-mer LDA- and frequency-based motifs. Motifs in topics 0 and 1 have IP values above 7, indicating they are primarily basic, while frequency-based motifs show a wider range, reflecting both basic and acidic characteristics. Frequency-based motifs also exhibit a higher GRAVY index than those from topic models, with topic 0 having a notably higher index than topic 1. Both topics demonstrate elevated hydrophobic moments, though topic 0 is less variable. In contrast, frequency-based motifs have lower hydrophobic moments, suggesting that topic model motifs are more amphipathic. Regarding helix and turn fractions, topics 0 and 1 show distinct ranges of helix fractions, while frequency-based motifs generally have higher helix fractions. Nonetheless, both types maintain similar trends in turn fractions, highlighting valuable insights for future research.

18-mer motifs: Figure 7 shows that LDA-derived motifs in topics 0-3 have higher values for IP, GRAVY, and HM compared to frequency-based motifs, indicating they are hydrophobic, amphipathic, and basic. In contrast, frequency-based motifs exhibit lower values for hydrophobicity, IP, and helix/turn fractions. Notably, topic 3 motifs display lower helix fraction and hydrophobicity, suggesting they can be both acidic and basic.

In summary, frequency-based motif analysis reveals how often motifs appear across sequences, providing a general overview of patterns. However, this method can be repetitive and may lack biological relevance due to the absence of context. The LDA topic model, on the other hand, captures semantic context and uncovers hidden structures, effectively grouping motifs with similar properties in a biological framework.

Comparison 2: amino acid compositions

Table 1 outlines the LDA-derived and frequency-based motifs for 4-mers, 14-mers, and 18-mers. To enhance our understanding of the amino acid compositions within these motifs, Figs. 8, 9, S7, and S8 (found in the SI) present both individual and grouped amino acid compositions. The grouped compositions categorize amino

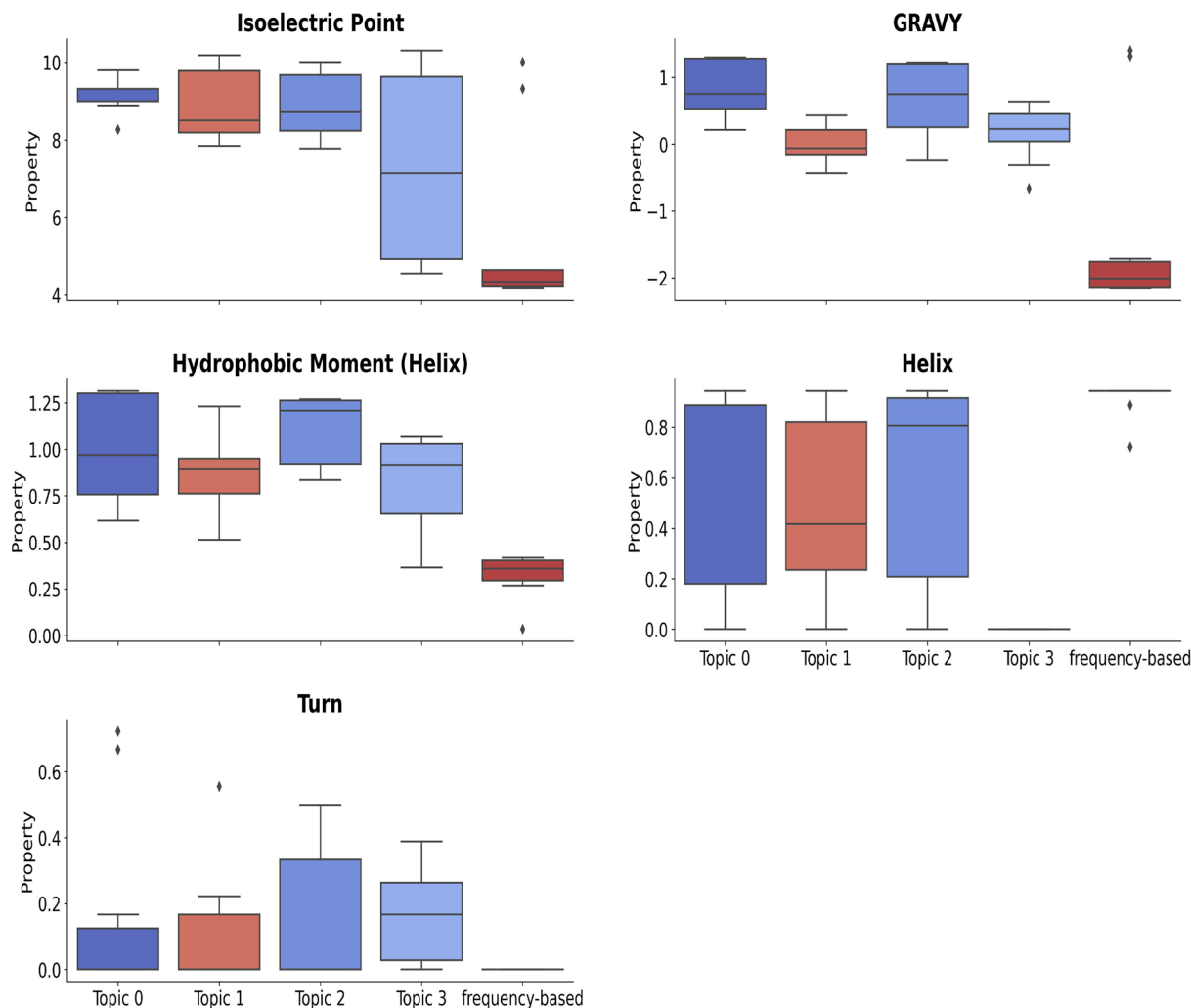


Fig. 7. The distributions of motif-level properties of LDA-derived and frequency-based 18-mer motifs. The LDA-derived motifs found in topics 0-3 exhibit higher values for IP, GRAVY, and HM compared to the frequency-based motifs. This suggests that the LDA-derived motifs are hydrophobic, amphipathic, and basic. In contrast, frequency-based motifs show lower values for hydrophobicity, IP, HTF. Notably, topic 3 motifs display lower values for helix fraction, hydrophobicity, and long-range isoelectric point.

acids as positively charged (K, R, H), negatively charged (D, E), hydrophobic (A, V, L, I, M, F, Y, W), polar uncharged (S, T, N, Q), and special cases (C, G, P). This concise analysis offers valuable insights into the motifs identified through the LDA topic model and frequency-based methods.

4-mer motifs: The analysis in Fig. 8 highlights the notable concentration of proline residues in LDA model-derived 4-mer motifs. These residues help maintain extended, open structures in peptides, facilitating their interaction with bacterial membranes. Proline-rich motifs, like "RPRP" often feature positively charged amino acids such as arginine, enabling effective engagement with negatively charged lipids in these membranes^{44,45}. This capability allows proline-rich AMPs to penetrate bacterial membranes without immediate cell lysis, reducing cytotoxicity while preserving antibacterial activity. Additionally, the motif compositions reveal a balance between charged and uncharged residues, which is critical for the hydrophobic properties of AMPs. Notably, Topic 1 (c & d) has a lower fraction of positively charged residues and a higher fraction of polar uncharged residues compared to Topic 0 (a & b).

14-mer motifs: Figure 9 shows the amino acid compositions of the top 10 14-mer motifs. Topic 1 (panels c & d) has a higher composition of positively charged residues but fewer special case residues, suggesting intriguing functional implications. In contrast, topic 0 (panels a & b) contains more glycine-rich motifs. Both topics display significant presence of cysteine and histidine residues, with cysteine being essential for disulfide bridge formation that stabilizes the beta-strand structure of AMPs⁴⁶⁻⁴⁸. Topic 0's high concentration of histidine residues provides pH-responsive cationic properties and facilitates metal binding, promoting membrane disruption and reactive oxygen species generation at acidic sites^{49,50}. Its structural roles enhance antimicrobial efficacy while minimizing host cell toxicity⁵¹⁻⁵³.

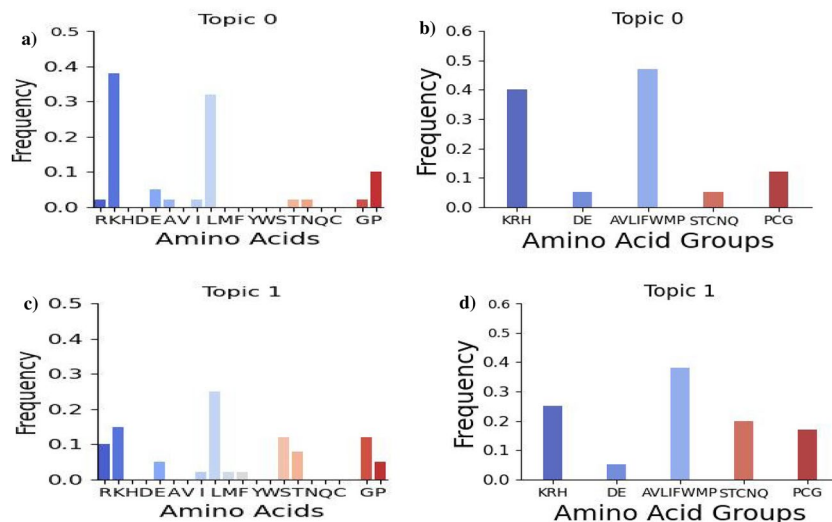


Fig. 8. Amino acid compositions for the top ten LDA-derived 4-mer motifs associated with the *E. coli* target. It depicts both individual amino acid compositions (a and c) for each of the top motifs and group-wise compositions (b and d) of topic 0 and 1. Specifically, it includes the amino acid compositions for positively charged amino acids (K, R, H), negatively charged amino acids (D, E), hydrophobic amino acids (A, V, L, I, M, F, Y, W), polar uncharged amino acids (S, T, N, Q), and special cases of amino acids (C, G, P).

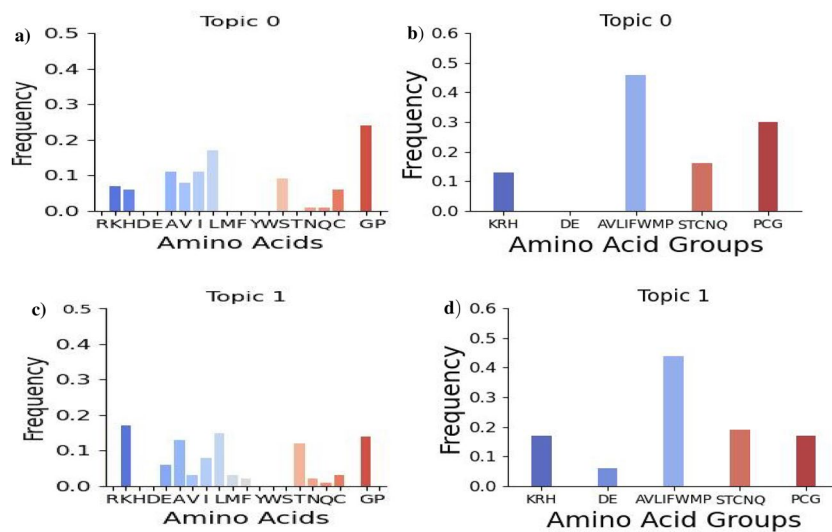


Fig. 9. Amino acid compositions for the top ten LDA-derived 14-mer motifs associated with the *E. coli* target. It depicts both individual amino acid compositions (panels a & c) for each of the top motifs and group-wise compositions (panels b and d) of topic 0 and 1.

18-mer motifs: Figure S7 presents an analysis of the amino acid compositions (Panels a, c, e, & g) and the group-wise aggregated compositions (b, d, f, & h) of 18-mer motifs across four topics. Topics 1 and 3 show enrichment in positively charged residues like lysine, which enhances antimicrobial activity through stronger interactions with negatively charged bacterial membranes. Conversely, Topics 0 and 2 are rich in uncharged residues, particularly serine (S), suggesting flexibility or potential phosphorylation sites. Glycine (G) is prevalent in all topics, highlighting its role in structural flexibility. In addition, hydrophobic residues are dominant across all topics, essential for membrane interactions. Notably, Topic 1 has a unique enrichment of threonine (T), which could influence motif polarity or function through hydrogen bonding. These variations indicate that motifs associated with each topic may serve distinct structural or functional roles, reflecting the underlying biological mechanisms relevant to AMP design.

Frequency-based motifs: Figure S8 shows the amino acid compositions for 4-, 14-, and 18-mer motifs from a frequency-based method (Panels a, c, & e) along with their group-wise compositions (b, d, & f). Shorter motifs (4- and 14-mers) are enriched in leucine and phenylalanine, leading to higher hydrophobicity. This characteristic aids in anchoring peptides in the bilayer through hydrophobic interactions, potentially causing membrane

thinning and the formation of defects or pores^{54,55}. In contrast, the 18-mer motifs exhibit greater diversity, with an increase in both positively and negatively charged residues. This balanced charge distribution may enhance functional complexity, supporting interactions such as forming amphipathic helices or engaging in electrostatic interactions with microbial membranes.

In summary, frequency-based motifs are rich in leucine and phenylalanine, especially in the 4- and 14-mer lengths, contributing to their strong hydrophobic properties. In contrast, LDA-derived motifs show lower phenylalanine levels but maintain hydrophobicity through alternative residues and exhibit increased lysine presence. This highlights the robust hydrophobic characteristics of frequency-based motifs while suggesting that the compositional diversity in LDA-derived motifs may lead to distinct antimicrobial functions.

Comparison 3: structural analysis

To better understand the motifs of the LDA topic model and frequency-based approaches, we analyzed their structural characteristics using the ESM fold model¹⁷. As shown in Fig. 10, the structural quality of 14-mers motifs was visualized using the pLDDT color map of ESMFold. Although ESMFold is not specifically optimized for short peptides, predicted models were used to qualitatively illustrate structural diversity between topics, with confidence levels indicated by the pLDDT scores. The motifs of topic 0 mainly exhibited random coil structures, while the last three showed some helical properties. In contrast, topic 1 motifs predominantly displayed helical structures, but the eighth and ninth motifs were exceptions, suggesting opportunities for further research into their structural capabilities.

Additionally, Fig. 11 indicates that topics 0 and 1 had lower MIC values for 14-mer sequences, highlighting Topic 1's potential as candidate AMPs. However, Frequency-based motifs demonstrated a more uniform structure. Figure S9 in the SI presents the structural analysis of the top ten LDA- and frequency-based 18-mer motifs. Here, topics 0-2 primarily exhibited helical structures, with only three showing random coils, while Topic 3 consisted entirely of random coils. This suggests distinct structural properties, and frequency-based motifs showed consistent structural uniformity, similar to the 14-mer motifs.

COMPARISON 4: sequence-level property (MIC)

This section compares LDA-derived motifs with frequency-based motifs through MIC sequence-level analysis. We assessed motif significance by examining their occurrences in actual sequences and correlating them with MIC values. This provides an approximation of motif potency but may not fully reflect the true activity of motifs in isolation. To complement this, in previous section we computed chemical properties at the motif level, enabling more reliable comparisons and interpretation of motif activity.

For instance, Fig. 11 shows that for 4-mer motifs, LDA-derived motifs in Topic 1 had lower MIC values than those in Topic 0, while frequency-based motifs also showed lower MIC values than Topic 0. For 14-mer motifs, frequency-based motifs exhibited higher MIC values than LDA-derived ones, with Topic 0 having the lowest

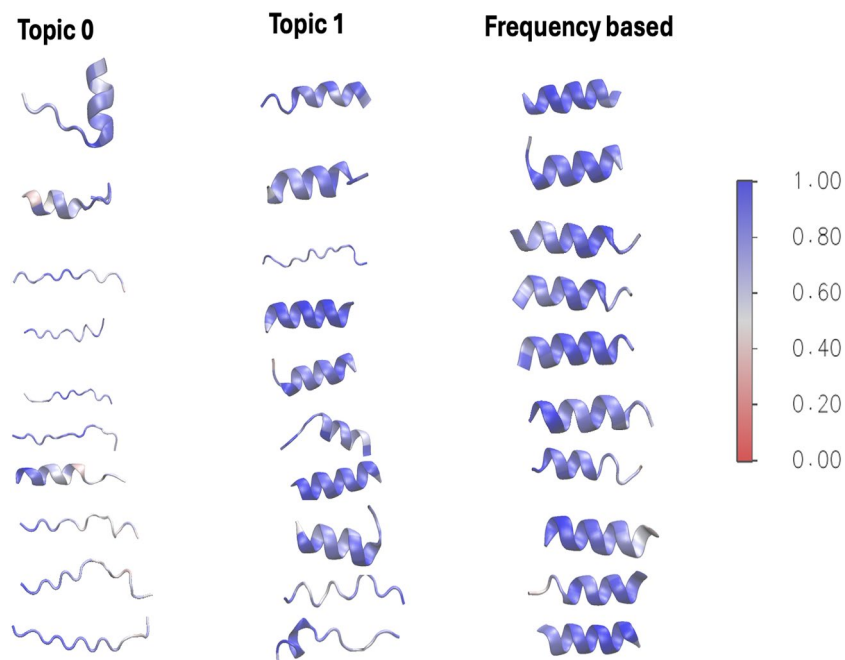


Fig. 10. The structural analysis of the top ten LDA- and frequency-based 14-mer motifs. Topic 0 primarily exhibit random coil structures, with the last three motifs displaying helices, while Topic 1 showcases helical structures, except for the 8th and 9th motifs. Frequency-based motifs show a more uniform structure overall, indicating they are less common than LDA-derived motifs in AMP analysis. Additionally, the color-bar indicates the confidence levels of the predicted structures as determined by the ESMFold model.

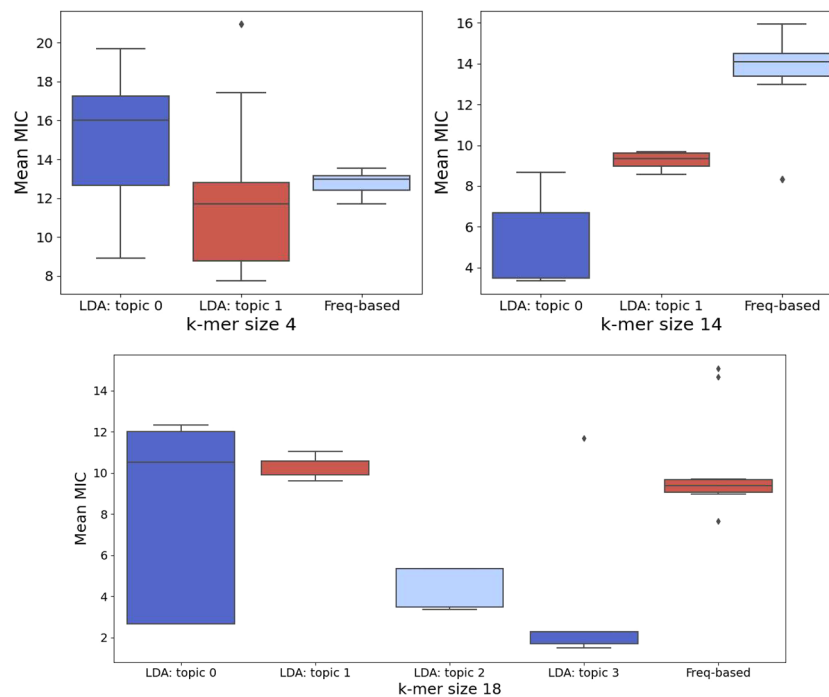


Fig. 11. Mean MIC values for LDA- and frequency-based motifs of lengths 4, 14, and 18. For 4-mers, LDA-derived motifs in topic 1 show lower MIC values than those in topic 0. In 14-mers, topics 0 and 1 have lower MIC values compared to frequency-based motifs. For 18-mer motifs, Topics 2 and 3 exhibit the lowest MIC values relative to topics 0 and 1, while frequency-based motifs consistently show higher MIC values, especially in topic 1.

values. In terms of 18-mer motifs, Topics 2 and 3 demonstrated the lowest MIC values, whereas frequency-based motifs correlated with higher MIC values in Topic 1. Overall, these findings suggest that motifs derived from the LDA model are more closely associated with sequences displaying lower MIC values.

Discussions

When using multiple databases, we have taken into account about redundant sequences by taking the average MIC value. However, highly similar sequences can lead to over representation of specific k-mers. Although frequency-based analyzes are affected by this redundancy, topic modeling uncovers hidden structures beyond high-frequency motifs. Our comparison of an LDA-based approach with a frequency-based method demonstrated that LDA effectively identifies biologically significant motifs, even in the presence of overlapping sequences across databases.

We evaluated the topic model by extracting shorter and longer motifs. Shorter motifs are particularly useful for generating diverse AMPs, while longer motifs capture richer contextual information within peptide sequences. For novel sequence design, we prioritize shorter motifs (e.g., 4-mers) to promote diversity, whereas longer motifs can be employed to design AMPs that are more similar to known targets. Thus, motif length serves as a design parameter that can be tuned depending on the desired balance between diversity and similarity when exploring new AMP candidates.

Frequency-based motif discovery demonstrates comparable MIC values at shorter motif lengths (4-mers), but LDA-derived motifs show significant advantages at longer lengths (14- and 18-mer motifs) in MIC association and diversity. This indicates that frequency alone may overlook essential long-range dependencies and contextual relevance crucial for understanding biological functions. Our findings highlight the effectiveness of topic modeling in revealing latent motif structures that carry functional insights. The LDA-derived motifs of topics 0 and 1 have MIC profiles similar to frequency-derived motifs at $k = 4$, but topics 0 and 1 present significantly lower MIC values ($p < 0.001$ in Table 3) for 14-mers, suggesting contextually relevant motifs missed by frequency-based methods.

Similarly, LDA-derived motifs of topics 2 and 3 yield lower MIC values for 18-mers. Mapping sequence MIC values to motifs and aggregating them based on occurrences can uncover connections between specific motifs and varying MIC levels, offering insights into AR. However, overlapping box plots may emerge when few sequences show outlier MIC values, complicating clear topic separations. The biological complexity also plays a role, as multiple factors influence the MIC, indicating that motifs alone may not fully account for these variations. Some biophysical properties of analyzed peptide motifs may not be entirely accurate, and our assumption of helical formation when calculating hydrophobic moments needs further validation.

Comparison of LDA motifs with those of other methods is challenging. LDA motifs are fixed-length sequences (e.g., 4, 14, or 18-mers), while deep learning⁵⁶ and attention-based models⁵⁷ produce longer variable-

K-mer size	Frequency-based (mean MIC)	Topic	LDA model (mean MIC)	p_value
4	12.73 ± 0.20	0	15.09 ± 1.13	0.95
		1	12.08 ± 1.34	0.05
14	13.61 ± 0.64	0	5.76 ± 0.67	< 0.001
		1	9.28 ± 0.13	< 0.001
18	10.26 ± 0.78	0	7.97 ± 1.46	0.45
		1	10.33 ± 0.16	0.98
		2	4.58 ± 0.31	< 0.001
		3	2.98 ± 0.97	< 0.001

Table 3. Shows the mean MIC for frequency-based and LDA-derived motifs computed using MIC sequence-level comparison, revealing that 14-mers from topics 0 and 1 had significantly lower MIC values (equal-variance t-test with $p_{\text{value}} < 0.001$). This suggests a high probability of rejecting the null hypothesis that the mean MIC of LDA-derived motifs is greater than that of frequency-based motifs.

length motifs. These models prioritize predictive accuracy over recurring sequences, whereas LDA identifies co-occurring fragments without supervision. The different methodologies make meaningful comparisons complex, but they are worth exploring in future research.

Conclusion and future work

We effectively demonstrated the LDA topic model's ability to uncover motifs linked to antimicrobial activity. Our method involves segmenting AMPs into k-mers, building an LDA model, and mapping motifs to relevant biological properties. Our results reveal that arginine (R)-rich sequences are particularly effective AMPs, while lysine (K) and leucine (L) motifs also show significant antimicrobial activity with lower MIC values. The LDA model successfully extracts contextually relevant motifs, which show lower MIC values as compared to frequency-based method. In conclusion, motifs are essential for identifying promising candidates for AMP design. By comparing the motifs present in low and high MIC groups, we see a clear correlation between specific motifs and antimicrobial effectiveness. These characteristics can be combined with various properties to enhance prediction and classification tasks, advancing our understanding of AMPs. In future work, we aim to utilize the identified motifs to design novel AMPs and evaluate their biochemical and biological properties, further exploring the significance of these motifs and the patterns revealed by our analysis.

Data availability

The code utilized for our experiments can be accessed at https://gitlab.com/padi.sarala-group/Topic_Model_Motif_Discovery. The folder contains scripts for topic model optimization, metrics, data analytics for biological outputs, pretrained weights for structure prediction using ESM fold, and feature extraction using Bio-Python.

Received: 18 September 2025; Accepted: 25 November 2025

Published online: 29 December 2025

References

- Craig, M. CDC's antibiotic resistance threats report, 2019. Extended spectrum β -lactamase (ESBL)-producing Enterobacteriaceae (2019).
- Yahn, E. Drug-resistant infections a threat to our economic future March 2017 2017 international bank for reconstruction and development/the world bank. *World* **135**, 256 (2016).
- Murray, C. J. et al. Global burden of bacterial antimicrobial resistance in 2019: a systematic analysis. *The lancet* **399**, 629–655 (2022).
- O'Neill, R. & Grande-Bretagne, J. Antimicrobial resistance: tackling a crisis for the health and wealth of nations on antimicrobial resistance (London). *Rev. Antimicrob. Resistance* (2014).
- World Health Organization. *Global Action Plan on Antimicrobial Resistance* (World Health Organization, 2015).
- Lopes, B. S. et al. The role of antimicrobial peptides as antimicrobial and antibiofilm agents in tackling the silent pandemic of antimicrobial resistance. *Molecules* **27**, 1452. <https://doi.org/10.3390/molecules27092995> (2022).
- Cardoso, P. et al. Molecular engineering of antimicrobial peptides: microbial targets, peptide motifs and translation opportunities. *Biophys. Rev.* **13**, 35–69 (2021).
- Xuan, J. et al. Antimicrobial peptides for combating drug-resistant bacterial infections. *Drug Resist. Updates* **68**, 100954 (2023).
- Moretta, A. et al. Antimicrobial peptides: a new hope in biomedical and pharmaceutical fields. *Front. Cell. Infect. Microbiol.* **11**, 668632 (2021).
- Mba, I. E. & Nweze, E. I. Focus: antimicrobial resistance: antimicrobial peptides therapy: an emerging alternative for treating drug-resistant bacteria. *Yale J. Biol. Med.* **95**, 445 (2022).
- Zasloff, M. Antimicrobial peptides of multicellular organisms. *Nature* **415**, 389–395 (2002).
- Wu, Q., Patočka, J. & Kuča, K. Insect antimicrobial peptides, a mini review. *Toxins* **10**, 461 (2018).
- Xiong, J. *Protein Motifs and Domain Prediction* 85–94 (Cambridge University Press, 2006).
- Liu, J. & Rost, B. Domains, motifs and clusters in the protein universe. *Curr. Opin. Chem. Biol.* **7**, 5–11 (2003).
- Mackenzie, C. O. & Grigoryan, G. Protein structural motifs in prediction and design. *Curr. Opin. Struct. Biol.* **44**, 161–167. <https://doi.org/10.1016/j.sbi.2017.03.012> (2017).
- Jacobs, T. et al. Design of structurally distinct proteins using strategies inspired by evolution. *Science* **352**, 687–690 (2016).
- Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).

18. Madani, A. et al. Large language models generate functional protein sequences across diverse families. *Nat. Biotechnol.* **41**, 1099–1106 (2023).
19. Elnaggar, A. et al. Prottrans: towards cracking the language of life's code through self-supervised learning. *bioRxiv* <https://doi.org/10.1101/2020.07.12.199554> (2021).
20. Jumper, J. et al. Highly accurate protein structure prediction with alphafold. *Nature* **596**, 583–589 (2021).
21. Liu, L., Tang, L., Dong, W., Yao, S. & Zhou, W. An overview of topic modeling and its current applications in bioinformatics. *Springerplus* **5**, 1–22 (2016).
22. Juan, L. et al. Evaluating individual genome similarity with a topic model. *Bioinformatics* **36**, 4757–4764 (2020).
23. Basha-Gutierrez, J. & Nakai, K. A study on the application of topic models to motif finding algorithms. *BMC Bioinform.* **17**, 129–138 (2016).
24. Schneider, N., Fechner, N., Landrum, G. A. & Stiefl, N. Chemical topic modeling: exploring molecular data sets using a common text-mining approach. *J. Chem. Inf. Model.* **57**, 1816–1831 (2017).
25. Li, C. et al. Amplify: attentive deep learning model for discovery of novel antimicrobial peptides effective against who priority pathogens. *BMC Genomics* **23**, 142. <https://doi.org/10.1186/s12864-022-08310-4> (2022).
26. Dang, T. Motifquest: an automated pipeline for motif database creation to improve database searching programs. *J. Am. Soc. Mass Spectrom.* **35**, 1902–1910. <https://doi.org/10.1021/jasms.4c00192> (2024).
27. Jensen, K. *Motif discovery in sequential data*. Ph.D. thesis, Massachusetts Institute of Technology (2006).
28. Castellana, S. A comparative benchmark of classic dna motif discovery tools on synthetic data. *Bioinformatics* **22**, 634–640. <https://doi.org/10.1093/bib/bbab303> (2021).
29. Vahed, M. Bml: a versatile web server for bipartite motif discovery. *Brief. Bioinform.* **23**, bbab536. <https://doi.org/10.1093/bib/bba536> (2022).
30. Gao, S. et al. Plptp: a motif-based interpretable deep learning framework based on protein language models for peptide toxicity prediction. *J. Mol. Biol.* **437**, 169115. <https://doi.org/10.1016/j.jmb.2025.169115> (2025).
31. Bhangui, S. K. et al. Machine learning-assisted prediction and generation of antimicrobial peptides. *Small Sci.* **5**, 2400579. <https://doi.org/10.1002/smcs.202400579> (2025).
32. Ofer, D., Brandes, N. & Linal, M. The language of proteins: Nlp, machine learning & protein sequences. *Comput. Struct. Biotechnol. J.* **19**, 1750–1758 (2021).
33. Blei, D. M. Probabilistic topic models. *Commun. ACM* **55**, 77–84. <https://doi.org/10.1145/2133806.2133826> (2012).
34. Řehůřek, R. & Sojka, P. Software framework for topic modelling with large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks* 45–50 (ELRA, 2010). <http://is.muni.cz/publication/884893/en>.
35. Cock, P. J. et al. Biopython: freely available python tools for computational molecular biology and bioinformatics. *Bioinformatics* **25**, 1422–1423 (2009).
36. Jacob, J., Duclouhier, H. & Cafiso, D. S. The role of proline and glycine in determining the backbone flexibility of a channel-forming peptide. *Biophys. J.* **76**, 1367–1376 (1999).
37. Witten, J. & Witten, Z. Deep learning regression model for antimicrobial peptide design. *BioRxiv* **2019**, 692681 (2019).
38. Wang, G., Li, X. & Wang, Z. Apd3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Res.* **44**, D1087–D1093 (2016).
39. 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.* **49**, D288–D297 (2021).
40. Piotto, S. P., Sessa, L., Concilio, S. & Iannelli, P. Yadamp: yet another database of antimicrobial peptides. *Int. J. Antimicrob. Agents* **39**, 346–351 (2012).
41. Fan, L. et al. Dramp: a comprehensive data repository of antimicrobial peptides. *Sci. Rep.* **6**, 24482 (2016).
42. Aguilera-Mendoza, L. et al. Graph-based data integration from bioactive peptide databases of pharmaceutical interest: toward an organized collection enabling visual network analysis. *Bioinformatics* **35**, 4739–4747 (2019).
43. Aguilera-Mendoza, L. et al. Starpep toolbox: an open-source software to assist chemical space analysis of bioactive peptides and their functions using complex networks. *Bioinformatics* **39**, btad506 (2023).
44. Scocchi, M., Tossi, A. & Gennaro, R. Proline-rich antimicrobial peptides: converging to a non-lytic mechanism of action. *Cell. Mol. Life Sci.* **68**, 2317–30. <https://doi.org/10.1007/s00018-011-0721-7> (2011).
45. Mishra, A. K., Choi, J., Moon, E. & Baek, K.-H. Tryptophan-rich and proline-rich antimicrobial peptides. *Molecules* **23**, 815 (2018).
46. Hocquellet, A., le Senechal, C. & Garbay, B. Importance of the disulfide bridges in the antibacterial activity of human hepcidin. *Peptides* **36**, 303–307 (2012).
47. Park, S.-H. et al. Role of proline, cysteine and a disulphide bridge in the structure and activity of the anti-microbial peptide gaegurin 5. *Biochem. J.* **368**, 171–182 (2002).
48. Datta, A., Kundu, P. & Bhunia, A. Designing potent antimicrobial peptides by disulphide linked dimerization and n-terminal lipidation to increase antimicrobial activity and membrane perturbation: Structural insights into lipopolysaccharide binding. *J. Colloid Interface Sci.* **461**, 335–345 (2016).
49. Hayashi, S. et al. Control of reactive oxygen species (ros) production through histidine kinases in aspergillus nidulans under different growth conditions. *FEBS Open Bio* **4**, 90–95 (2014).
50. Paretzoglou, A., Stockenhuber, C., Kirk, S. & Ahmad, S. Generation of reactive oxygen species from the photolysis of histidine by near-ultraviolet light: effects on t7 as a model biological system. *J. Photochem. Photobiol. B* **43**, 101–105 (1998).
51. McDonald, M. et al. Structure–function relationships in histidine-rich antimicrobial peptides from atlantic cod. *Biochim. Biophys. Acta (BBA) Biomembranes* **1848**, 1451–1461 (2015).
52. Kacprzyk, L. et al. Antimicrobial activity of histidine-rich peptides is dependent on acidic conditions. *Biochim. Biophys. Acta (BBA)-Biomembranes* **1768**, 2667–2680 (2007).
53. Amirkhanov, N., Bardasheva, A., Tikunova, N. & Pyshnyi, D. Synthetic antimicrobial peptides: Iii–effect of cationic groups of lysine, arginine, and histidine on antimicrobial activity of peptides with a linear type of amphipathicity. *Russ. J. Bioorg. Chem.* **47**, 681–690 (2021).
54. Epan, R. M. & Vogel, H. J. Diversity of antimicrobial peptides and their mechanisms of action. *Biochim. Biophys. Acta (BBA)-Biomembranes* **1462**, 11–28 (1999).
55. Chen, Y. et al. Role of peptide hydrophobicity in the mechanism of action of α -helical antimicrobial peptides. *Antimicrob. Agents Chemother.* **51**, 1398–1406 (2007).
56. Wang, B. et al. Explainable deep learning and virtual evolution identifies antimicrobial peptides with activity against multidrug-resistant human pathogens. *Nat. Microbiol.* **10**, 332–347. <https://doi.org/10.1038/s41564-024-01907-3> (2025).
57. Li, C., Zou, Q., Jia, C. & Zheng, J. Ampred-mfa: an interpretable antimicrobial peptide predictor with a stacking architecture, multiple features, and multihead attention. *J. Chem. Inf. Model.* **64**, 2393–2404. <https://doi.org/10.1021/acs.jcim.3c01017> (2024).

Disclaimer

The commercial products used in this study were only referenced to specify the experimental procedure adequately. Such identification of commercial products is not intended to imply recommendation or endorsement by the National Institute of Standards and Technology, nor is it intended to imply that the identified products are necessarily the best available for the purpose.

Author contributions

SP, AC, DH, FH, and MM conceived the study. SP developed the methodology, was responsible for experimental evaluations, carried out formal analysis and investigated. SP and KM curated data. SP wrote the original article draft, which was reviewed by DH, FH, MM, JBK and AC. SP was responsible for visualizations, KM was responsible for motif-level property extraction and visualizations. MM, FH, DH, JBK, and AC supervised this study.

Funding

This work was supported by an Innovation in Measurement Science (IMS) grant from the National Institute of Standards and Technology (70NANB17H299 and 70NANB24H248).

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1038/s41598-025-30419-1>.

Correspondence and requests for materials should be addressed to S.P.

Reprints and permissions information is available at www.nature.com/reprints.

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

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

This is a U.S. Government work and not under copyright protection in the US; foreign copyright protection may apply 2025