bioRxiv preprint doi: https://doi.org/10.1101/2025.03.19.644165; this version posted March 19, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The Use of DeepQSAR Models for The Discovery of Peptides with Enhanced Antimicrobial and Antibiofilm Potential.

Jiaying You (1), Evan Haney (2), Noushin Akhoundsadegh (2), Robert E.W. Hancock (2), Artem Cherkasov (1*)

1) Vancouver Prostate Centre, Department of Urologic Sciences, Faculty of Medicine, University of British Columbia, Vancouver, BC, Canada 2) Centre for Microbial Diseases and Immunity Research, Department of Microbiology and Immunology, Faculty of Science, University of British Columbia, Vancouver, BC, Canada

**** Correspondence: acherkasov@prostatecentre.com***

Abstract

Increasing concerns regarding prolonged antibiotic usage have spurred the search for alternative treatments. Antimicrobial peptides (AMPs), first discovered in the 1980s, have exhibited significant potential against a broad range of bacteria [1]. Short-sequenced AMPs are abundant in nature and present across various organisms [2]. With lower likelihood of developing resistance, thousands of cationic and hydrophobic AMPs have been identified to act selectively on microbes, making them excellent anti-infective candidates in medicine, food, farming, and agriculture. Recently, machine learning technologies such as Quantitative Structure Activity Relationships (QSAR) have enabled expedited discovery of potential AMPs with broad-spectrum antibacterial activity as the amount of available AMP training data increases [3]. Among those, Deep QSAR has recently emerged as a distinct type of application that utilizes conventional molecular descriptors [4] in conjunction with more powerful deep learning (DL) models [5]. Here, we demonstrate the power of Deep QSAR in predicting broad-spectrum AMP activity. Our approach utilizes a recurrent neural network to perform QSAR regression

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.19.644165; this version posted March 19, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

modeling and achieves nearly 90% five-fold cross-validated accuracy in predicting AMP antimicrobial and antibiofilm potential. Using the developed approach, we designed 100 novel peptides, of which 44 experimentally demonstrated more effective antibiofilm activity and 31 peptides exhibited stronger antimicrobial activity compared to the known antibiofilm/antimicrobial peptide IDR-1018 [6] used as a positive control. Notably, 29 of those peptides demonstrated improvements of both antimicrobial and, particularly, antibiofilm properties, making them suitable prototypes for pre-clinical development and demonstrating efficacy of DeepQSAR modeling in identifying novel and potent AMPs.

Introduction

Antibiotic pollution is becoming a global environmental challenge, driven by everyday human activities such as livestock farming, agriculture, waste disposal and urban wastewater discharge [7]. The inappropriate disposal of antibiotics to the air, soil, and water has forced living organisms to develop defensive mechanisms, namely antimicrobial resistance (AMR) [8], weakening the effectiveness of modern medicine in treating infections. As a result, healthcare costs have risen, as additional treatments and interventions become necessary to combat resistant infections.

One promising alternative to the conventional use of small molecule antibiotics is to utilize antimicrobial peptides (AMPs). These naturally occurring substances are found in nearly all living organisms and serve as a crucial component of the immune system, providing broad-spectrum protection against bacteria, fungi, parasites, and viruses and biofilms and modulating immunity. Notably, AMPs exhibit unique mechanisms of potent, selective and direct bacterial destruction resulting in a lower propensity for pathogens to develop resistance [9,10].

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.19.644165; this version posted March 19, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Overall, the growing concern regarding AMR has driven intensive discovery and design of novel AMPs with enhanced efficacy and stability. In recent years, advances in computational tools [11] have significantly accelerated the process in identifying novel potent peptides. For example, machine learning (ML) [12–14] and deep learning (DL) [15–17] models have been increasingly employed to analyze large biological datasets, to predict AMP and antibiofilm activity, and to generate novel peptide sequences with desired properties. These approaches offer a more efficient and cost-effective strategy for identifying potential antimicrobial peptides, reducing the reliance on traditional experimental methods. As research in this field progresses, computationall- designed AMPs could play an important role in addressing the global challenge of antimicrobial resistance and developing next-generation treatments.

ML [18] is a subset of artificial intelligence (AI) [19] that enables systems to learn from data, identify patterns, and make decisions without explicit programming. ML algorithms improve their performance over time by analyzing historical data and adjusting their model accordingly. ML is broadly classified into supervised learning [20] and unsupervised learning [21], and is widely used in applications such as fraud detection [22], recommendation systems [23] and biomarker prediction [24], driving automation [25] across different industries. In 2018, Yoshida et al. utilized machine learning combined with a generic algorithm and designed an evolutionary method for novel peptide generation. Without prior knowledge or databases, their method significantly enhanced antimicrobial activity, achieving a 162-fold improvement within three iterations, demonstrating that ML predictions efficiently navigate the vast peptide sequence space, reducing the need for extensive chemical synthesis and accelerating AMP discovery [26].

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.19.644165; this version posted March 19, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

DL is a specialized branch of machine learning that utilizes artificial neural networks with multiple hidden layers to process complex data [27]. DL architecture, such recurrent neural networks (RNNs) [28], is designed to mimic human cognitive functions, allowing them to learn hierarchical representations of data. Long short-term memory (LTSM) networks, a subtype of RNN neural network, are designed to effectively learn and process sequential data, which makes them promising for peptide design. Research has been conducted using LTSM models to generate new peptide sequences [29,30,31]. LSTM overcome the limitation due to the vanishing gradient [32] problem produced by RNN, by incorporating a unique gating mechanism that can regulate information flow, allowing them to maintain long-range dependencies and selectively update, ignore, or pass on the information as needed. In 2022, Li et.al [33] utilized multi-branch CNNs, where multiple parallel CNN structures were implemented, each followed by a semantic learning component to better capture peptide features. Additionally, they replace the gated recurrent unit (GRU) of RNNs with a bidirectional long short-term memory (Bi-LSTM) network, which improves the model's ability to capture contextual dependencies in peptide sequences. These designs were used here to develop more effective representations of bioactive peptides, ultimately achieving high prediction accuracy.

Results

It should be noted that research on simultaneous prediction of AMP multiple endpoints such as antimicrobial and antibiofilm activities of peptides remain limited, necessitating the development of a more comprehensive predictive model. To address this gap, we designed a combined Deep QSAR model that integrates regression-based numerical predictions with a sequence-based LSTM model using transfer learning.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.19.644165; this version posted March 19, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

This combined QSAR model enhanced the accuracy and efficiency of antimicrobial peptide (AMP) discovery by integrating advanced ML techniques. Our approach involved two key models: a regression model (Model 1) and a recurrent neural network (RNN) model (Model 2). Model 1 was designed to predict antibiofilm IC_{50} values of peptides using an in-house dataset with one-hot encoded peptide sequences, enabling numerical predictions of biofilm inhibition efficacy. This regression model served as a foundational step, capturing quantitative relationships between peptide sequences and their anti-biofilm activity.

Building upon the insights gained from Model 1, we constructed Model 2, an RNN-based binary LSTM classification model, to determine whether a given unknown peptide is likely to exhibit antimicrobial properties. To improve the RNN's predictive capability, we transferred the learned weights from Model 1, allowing it to leverage the peptide characteristics derived from the regression model. The integration process involved concatenating the output of Model 1 with the RNN architecture of Model 2, enabling the classification model to utilize numerical insights, while focusing on sequence-based patterns essential for antimicrobial prediction. This combined approach facilitated a more effective learning process, enhancing model performance and improving the accuracy of AMP identification. Through the use of this method, we demonstrated the potential of transfer learning and sequence-based features in advancing computational peptide discovery. To further validate our approach, we conducted a large-scale screening process followed by wet lab experiments.

In particular, we extracted 20,417 reviewed human protein sequences from Uniprot [34], filtered out proteins shorter than 100 amino acids, and systematically divided them into predicted peptide fragments, each 12 amino acids in length, using a sliding window approach. Novel peptide

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.19.644165; this version posted March 19, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

generation. Using the developed combined QSAR model, we then screened approximately 50,000 peptides to predict their antimicrobial and antibiofilm activities. Based on the model's positive predictions, we clustered the peptides into 100 unique clusters to ensure selection diversity, then selected top-ranked peptides in each cluster for chemical synthesis and subsequent biological testing. The 100 synthesized peptides were experimentally evaluated for both biofilm inhibition and planktonic antimicrobial activity using peptide arrays. Their effectiveness was assessed compared a known antibiofilm and antimicrobial peptide IDR-1018, used as a positive control. Among the 100 tested peptides, 44 peptides demonstrated more effective antibiofilm activity than the control peptide IDR-1018, 31 peptides exhibited stronger antimicrobial activity, 29 exhibited superior inhibition of both biofilm formation and planktonic bacterial growth compared to the positive control, confirming the model's ability to identify potent antimicrobial peptides. These experimental results provided strong validation of our computational approach, demonstrating that our combined model can effectively predict bioactive peptides with multi-functional antimicrobial properties.

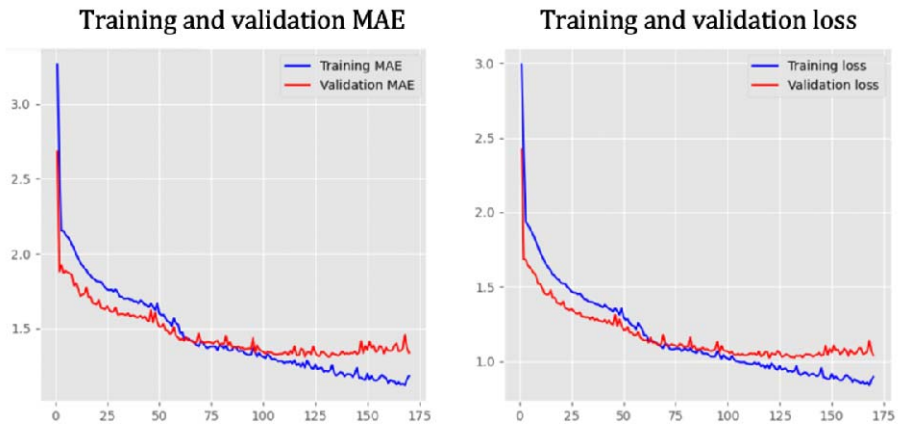
Model performance

The developed model demonstrated strong performance throughout the training and validation phases, as reflected in key evaluation metrics. Thus, the training and validation mean absolute error (MAE) [35] and loss curves indicate stable learning progress, with decreasing trends over time, suggesting that both Model 1 and Model 2 successfully captured meaningful patterns in the data [Figure 1]. The precision-recall curve highlighted the model's high precision across different recall levels, while the receiver operating characteristic (ROC) curve showed a high true positive rate with a low false positive rate, reinforcing the reliability of our predictions [Figure 2].

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.19.644165; this version posted March 19, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

To quantitatively assess the performance of the combined model, we evaluated precision, recall, and F1 score. The combined model achieved a precision of 0.90 for positive samples and 0.87 for negative samples, with recall values of 0.88 and 0.89, respectively. The F1 scores of 0.89 for both classes indicated a well-balanced performance in predicting antimicrobial peptides. These results indicated that the developed combined approach was highly effective, outperforming most existing AMP prediction models and demonstrating strong generalization in predicting unseen peptide sequences.

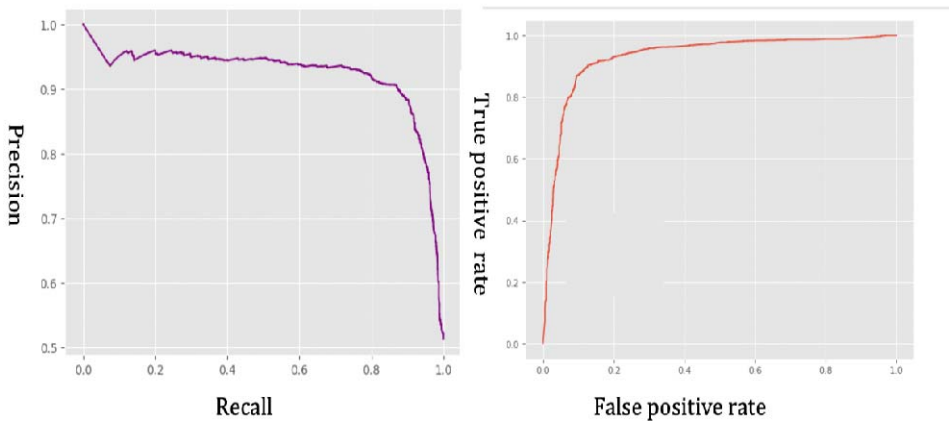
FIGURE 1: Regression model performance on training and



validation mean absolute error (MAE) and loss for Model 1.

FIGURE 2: precision-recall (left) and ROC (right) curves for Model 2.

Peptides synthesis and validation



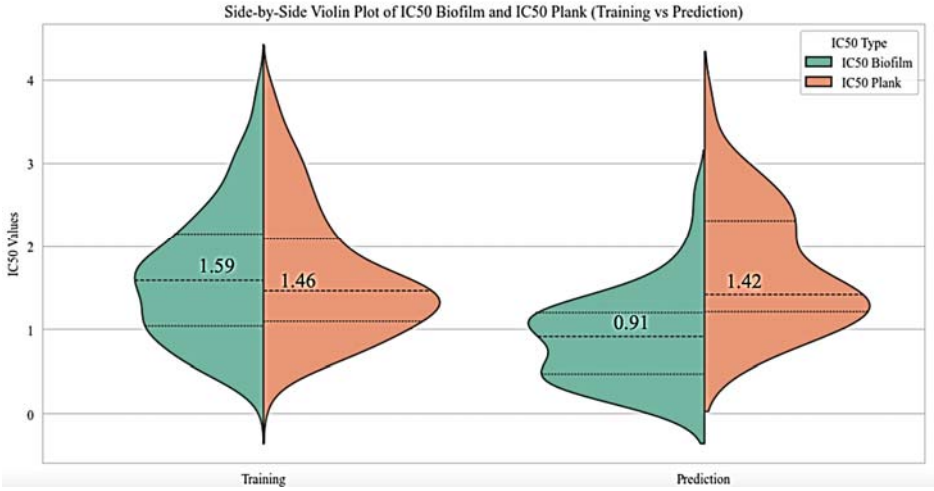
bioRxiv preprint doi: https://doi.org/10.1101/2025.03.19.644165; this version posted March 19, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The antimicrobial activity and biofilm inhibition activity of peptide array synthesized samples were assessed against the important AMR human pathogen methicillin resistant *Staphylococcus aureus* (MRSA) [36]. Based on model predictions and clustering, we selected the top 100 peptides for synthesis and experimental validation. Their antimicrobial efficacy was tested against both biofilm and planktonic bacterial populations, with activity compared to a known antimicrobial peptide (AMP), IDR-1018 [6], which served as a control. The results provided insights into the relative potency of these peptides in inhibiting bacterial growth and biofilm.

Among tested peptides, several exhibited superior anti-biofilm and anti-planktonic activities compared to the control, as measured by IC50 values (μM). Overall, we compared the IC_{50} values vs. Biofilms and planktonic bacteria between the peptides used in model training and the predicted novel peptides. The data for both sets were visualized using side-by-side violin plots [Figure 3], which illustrate the distribution of IC_{50} values for each group. The median IC_{50} values for these peptides in both antibiofilm and antimicrobial assays were consistently lower than those for the training peptides, providing strong evidence that the model can identify novel peptides with greater antibiofilm (and antimicrobial) activity.

FIGURE 3: IC_{50} values comparison on antibiofilm and antimicrobial activities between training samples used in Model 1 and synthesized novel peptides. Median values are shown on the charts

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.19.644165; this version posted March 19, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



Specifically, 44 peptides demonstrated more effective antibiofilm activity than the control peptide IDR-1018, 31 peptides exhibited stronger antimicrobial activity, and 29 peptides outperformed the control for both activities [Figure 4]. The top 10 peptides for anti-biofilm and anti-microbial activity are listed in the Tables 1 and Table 2. Generally, peptides performed better against biofilms, a top target for such peptides and one for which conventional agents are poorly available. Notably, 5 peptides were common to both top 10 lists, highlighting their exceptional efficacy in anti-biofilm and anti-microbial performance [Table 3, Figure 5].

FIGURE 4: Effectiveness of synthesized peptides.

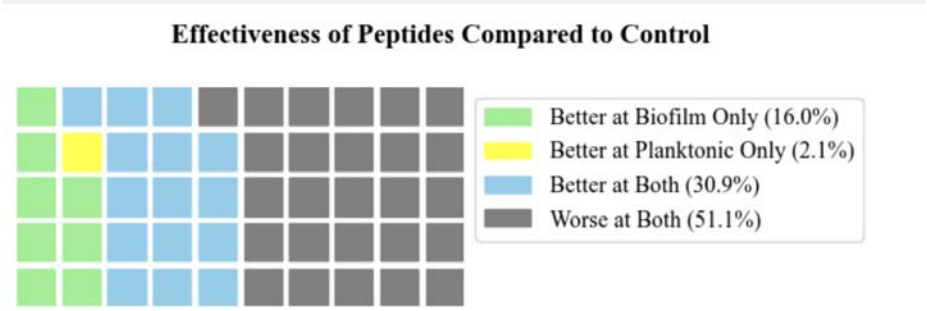
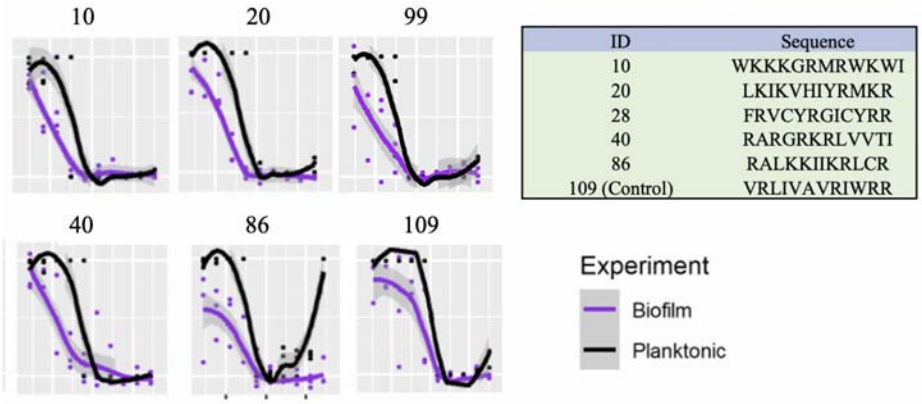


FIGURE 5: Array peptide biofilm vs. MRSA

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.19.644165; this version posted March 19, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.



ID	Sequence
10	WKKKGRMRWKWI
20	LKIKVHIYRMKR
28	FRVCYRGICYRR
40	RARGRKRLVVTI
86	RALKKIIKRLCR
109 (Control)	VRLIVAVRIWRR

Notably, peptide MVLRIKLRLKIR (ID 105) demonstrated the strongest biofilm inhibition with an IC_{50} of 0.147 μM, significantly lower by nearly 10-fold than the prominent anti-biofilm peptide positive control, IDR-1018 (1.417 μM). Additionally, it demonstrated improved planktonic activity (IC50 = 1.29 μM) relative to the control (1.73 μM). Other peptides, such as RALKKIIKRLCR (ID 86) and RARGRKRLVVTI (ID 40), also exhibited enhanced biofilm inhibition with IC_{50} values of 0.38 μM and 0.30 μM, respectively.

TABLE 1: Top 10 effective antibiofilm peptides with IC_{50} value (in µM) on biofilms compared with the positive control peptide IDR-1018 [6]. All peptides were amidated at the C-terminus.

ID	Sequence	IC_{50} Biofilm
105	MVLRIKLRLKIR	0.15
39	RGFVRLKKWFNI	0.23
99	MLIRVRKLWRIL	0.24
10	WKKKGRMRWKWI	0.27
59	FRVCYRGICYRK	0.30
40	RARGRKRLVVTI	0.30
28	FRVCYRGICYRR	0.35
20	LKIKVHIYRMKR	0.35
86	RALKKIIKRLCR	0.38
IDR-1018	VRLIVAVRIWRR	1.42

TABLE 2: Top 10 effective antimicrobial peptides with IC_{50} value (in µM)

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.19.644165; this version posted March 19, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

for planktonic (medium-based) organisms compared with the positive control IDR-1018 [6]. All peptides were amidated at the C-terminus.

ID	Sequence	IC_{50} antimicrobial
99	MLIRVRKLWRIL	0.70
86	RALKKIIKRLCR	0.71
10	WKKKGRMRWKWI	0.74
102	VLRIGWILWRIS	0.84
62	RRRAKGRIRLIV	0.89
100	LLILWRKLWILR	1.02
2	GRMRWKWIKKRI	1.03
20	LKIKVHIYRMKR	1.07
33	GLKSFARVLKKI	1.15
40	RARGRKRLVVTI	1.18
IDR-1018	VRLIVAVRIWRR	1.73

TABLE 3: Top 5 peptides extracted from top 10 rankings: dual superior anti-biofilm and anti-microbial activity based on IC_{50} values compared to the positive control peptide IDR-1018 [6]. All peptides were amidated at the C-terminus.

ID	Sequence	IC_{50} Biofilm	IC_{50} Plank
10	WKKKGRMRWKWI	0.27	0.74
20	LKIKVHIYRMKR	0.35	1.07
99	MLIRVRKLWRIL	0.24	0.70
40	RARGRKRLVVTI	0.30	1.18
86	RALKKIIKRLCR	0.38	0.70
IDR-1018	VRLIVAVRIWRR	1.42	1.73

These findings highlight the effectiveness of our combined Deep QSAR model in identifying peptides with superior antibiofilm and antimicrobial properties, particularly in inhibiting biofilm formation, which is crucial for combating chronic bacterial infections, a major target of such peptides[41] and one for which ML approaches are rare.

Methods

In-House Training Set

Peptide arrays on cellulose membranes were made by Kinexus

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.19.644165; this version posted March 19, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Bioinformatics Corporation (Vancouver, BC, Canada) and were obtained as cleaved free peptide associated with the punched-out spot of the cellulose membrane. Stock solutions of the peptide array samples were prepared by dissolving the free peptide in 200 μl of endotoxin free sterile water, incubating at room temperature for ~1 hr., and then transferring the resulting peptide solution to a sterile microfuge tube for subsequent antimicrobial activity assays. The concentration of the stock peptide array samples was determined by measuring the absorbance at 280 nm of the solution (in duplicate) on a BioTek Epoch Microplate Spectrophotometer using the Take3TM Multi-Volume Plate. A theoretical extinction coefficient of each peptide was determined based on the number of Trp and Try residues1 in the peptide sequence according to the formula: (#Trp) * 5500 M-1cm-1 + (#Tyr) * 1490 M-1cm-1, and the concentration of the peptide in solution was calculated according to the Beer-Lambert law. Any Cys residues within the peptide samples were assumed not to be present in a disulphide bond, so their contribution to the extinction coefficient at 280 nm was ignored. Stock concentrations of peptides lacking a Trp and Tyr residue were assigned the average concentration of all the samples for which a concentration could be determined, which was 393 μM.

This dataset comprised approximately 700 unique peptides, each tested for anti-biofilm IC_{50} values, resulting in 3,000 series arrays of numerical data. The IC_{50} values quantify the inhibitory concentration required to reduce biofilm formation by 50%, serving as a key indicator of peptide efficacy.

Antimicrobial activity

Peptides were synthesized on peptide arrays and tested as described previously [36]. Briefly, the antimicrobial activity and biofilm inhibition

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.19.644165; this version posted March 19, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

activity of the peptide array samples was assessed against methicillin resistant *Staphylococcus aureus* (MRSA) [36]. Briefly, two-fold serial dilutions of the peptide array sample (10 µl final volume) were prepared in 96-well plates and then 90 µl of a 1:100 dilution of an overnight culture of MRSA diluted in 10% tryptic soy broth supplemented with 0.1% glucose was added to each well and the plates were incubated overnight at 37◦C. The following day, bacterial growth was quantified by measuring the optical density at 600 nm (OD_{600}) of each sample well while the amount of biofilm was quantified by rinsing the wells of the plate with distilled water and staining the adhered biomass with 0.1% crystal violet solution. The amount of adhered biomass was quantified by recording the absorbance at 595 nm (A_{595}). Dose response curves were drawn for each peptide based on the bacterial growth (OD_{600}) or biofilm inhibition (A_{595}) and normalizing the percent of peptide inhibition relative to untreated bacteria (100% growth) or sterile media (0% growth). The concentration of peptide required to inhibit 50% of bacterial or biofilm growth (IC_{50}) was calculated by non-linear regression fit of the dose-response curves using the N-Parameter Logistic Regression (nplr) package.

Public Training Datasets

Three publicly available antimicrobial peptide (AMP) datasets were incorporated.

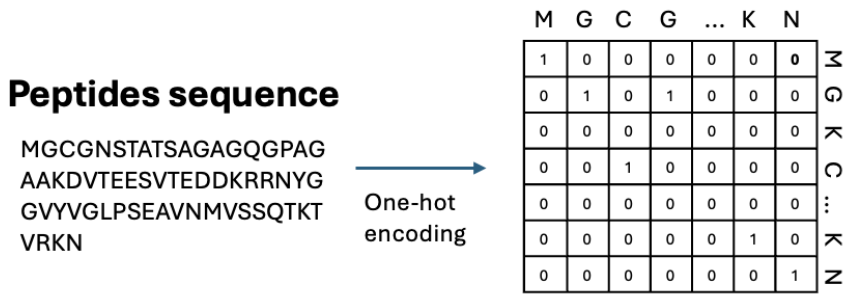
1. DRAMP (Database of Antimicrobial Peptides) [37]. Contains 22,259 peptide entries.
2. AI4AMP [38]. A balanced dataset consisting of 10,716 positive AMPs and 10,718 negative AMPs.
3. DBAASP [39] (Database of Antimicrobial Activity and Structure of Peptides). Contains 19,751 active AMPs with experimentally validated antimicrobial properties.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.19.644165; this version posted March 19, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The combination of these datasets ensured the model learnt from a diverse range of peptide sequences for classification model, enabling better generalization across different peptide properties and bioactivity profiles.

One-Hot Encoding of Peptide Sequences

One-hot encoding is a technique used to convert categorical data, such as peptide sequences, into a format that can be efficiently processed by



machine learning models [40]. In the context of peptide sequences, one-hot encoding transforms each amino acid into a unique binary vector, allowing machine learning algorithms to process the sequence data while preserving the order of the amino acids. We encoded 20 standard amino acids to unique binary vectors, each of length 20 [Figure 6]. These vectors are embedded where ’1’ at the position corresponds to the amino acid in the set, and ’0’s in all other positions. The entire peptide sequence is encoded as a matrix, where each row corresponds to the one-hot vector representation of the amino acids in the peptide. The of the key features of one-hot encoding is that it maintains the order of amino acids in the sequence.

F**IGURE 6: One-hot encoding of peptides sequences:** Each amino acid is represented as a unique binary vector while preserving sequence order.

Deep QSAR Modeling

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.19.644165; this version posted March 19, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Model 1: Regression.

We began by constructing a regression model (Model 1) using the in-house dataset, with IC_{50} values on anti-biofilm activities. The regression model was trained on one-hot encoded input vectors, focusing on numeric predictions of biofilm inhibition efficacy. To ensure a robust baseline, the model learned the quantitative relationships between the sequence features of peptides and their corresponding anti-biofilm activity. The dataset used for this model consisted of peptides with known biofilm inhibitory properties, and the model performance was evaluated by comparing its predicted IC_{50} values to experimentally determined values. This regression model provided a solid foundation for further development and enhanced our understanding of the quantitative aspects of biofilm inhibition.

Model 2: Recurrent Neural Network (RNN)

To enhance the performance of Model 2, we transferred the learned weights from Model 1 into the RNN architecture. This transfer of weights allowed the RNN to leverage the numerical insights and patterns learned from the regression model, facilitating the classification task. The RNN was then trained on a labeled dataset of peptides, each with a known antimicrobial property. These labels allowed the RNN to focus on learning sequence-specific patterns that contribute to antimicrobial activity, while also benefiting from the regression model's predictions of biofilm inhibition.

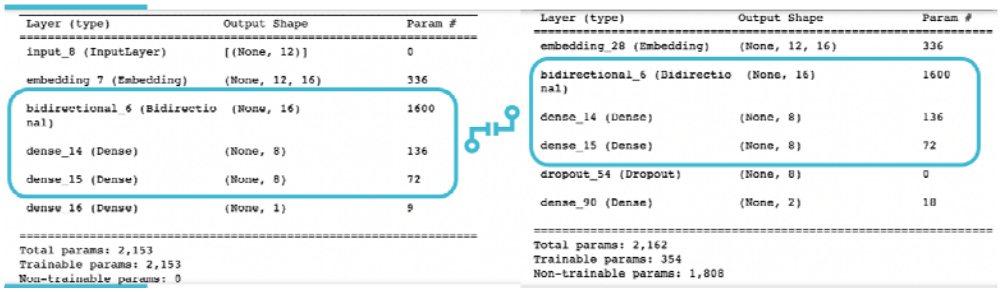
Combined Approach Integrating both Model 1 and Model 2

The integration of Model 1 and Model 2 was carried out by concatenating the output of the regression model (Model 1) with the RNN architecture of Model 2 [Figure 7]. This combined approach provided the RNN with valuable numerical insights from Model 1, which were incorporated as

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.19.644165; this version posted March 19, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

additional features to assist in classification. By using the regression model's predictions as input, the RNN was able to focus not only on sequence-based patterns but also on the broader quantitative relationships between peptide structure and biofilm inhibition, which were learned from the regression model.

FIGURE 7: Model architecture. Model 1 is a regression model built on IC_{50} values on anti-biofilm activities. Model 2 is a classification model on anti-microbial property with weights learnt from model 1.



The transfer of weights from Model 1 to Model 2 ensured a smoother transition of learned features and improved the RNN's performance. The RNN's ability to predict antimicrobial activity was enhanced by the foundation provided by the regression model, creating a more effective classifier for identifying antimicrobial peptides.

This integrated approach, combining the quantitative insights of the regression model with the sequential learning ability of the RNN, allowed for more accurate predictions of peptide efficacy, especially in terms of both biofilm inhibition and conventional antimicrobial properties.

Peptide Clustering

We applied a clustering method for narrowing down the number of peptides to be synthesized. To keep the diversity of peptide selection, the clustering method unitized the pairwise peptides sequences. We first calculated peptide similarity through global alignment, by aligning every pair of sequences from the beginning to end of each peptide. Then a

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.19.644165; this version posted March 19, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

symmetric similarity matrix was generated as the input to the hierarchical clustering approach (AgglomerativeClustering from sklearn.cluster), which uses Euclidean distance to complete the linkage between clusters. The highest prediction score peptide from our combined QSAR model was selected from each cluster, ensuring broad representation from all regions of peptide sequence space.

Discussion

Here, we developed a Deep QSAR model capable of identification of novel enhanced anti-biofilm and antimicrobial peptides. Through the filtering of human protein sequences in the UniProt database and selection of peptide fragments, we filtered 50,000 potential candidates and identified peptides with experimentally-confirmed enhanced activity when compared to the positive control anti-biofilm, antimicrobial peptide IDR-1018 [6].

Twenty-nine of the predicted peptides exhibited significantly lower IC_{50} values for biofilm and bacterial inhibition compared to the control. For instance, WKKKGRMRWKWI (ID 10) exhibited a 5.3-fold higher biofilm inhibition (IC50 = 0.266 μM) and a 2.3-fold higher antibiotic-like inhibition (IC50 = 0.745 μM) compared to IDR-1018. Other peptides, such as RARGRKRLVVTI (ID 40) and RALKKIIKRLCR (ID 86), also exhibited enhanced biofilm inhibitory activity. Overall, 46% of the top peptides reported higher biofilm inhibition than the excellent control peptide 1018, 32% exhibited greater antimicrobial (antibiotic) activities, and nearly 30% of peptides exhibited improved properties in both cf. IDR-1018. These findings indicate the potential of our model in optimizing peptide discovery for antimicrobial applications.

One of the strongest findings of our approach is the integration of both sequence-based screening and predictive modeling, which allowed us to

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.19.644165; this version posted March 19, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

successfully narrow down a high-throughput dataset to a highly active subset of peptides. Unlike conventional AMP discovery methods requiring trial-and-error screening, our approach enhances the likelihood of identifying functional peptides with improved bioactivity.

Despite such promising results, further investigations are needed to establish the peptides' mechanism of action, spectrum of activity and stability, toxicity, and activity in in vivo models. Additionally, structural analysis and molecular dynamics simulations may also provide deeper insights into the interactions between these peptides and bacterial membranes, guiding further optimization.

Overall, our findings demonstrate the value of computational models in antimicrobial peptide discovery and uncover the potential for the development of novel therapeutics to combat biofilm-associated and resistant infections. Future research should focus on refining the model and expanding its applications to identify peptides with even broader antimicrobial profiles.

Conclusions

This study demonstrates significant enhancement in identifying novel antibiofilm and antibiotic-like peptides through a deep learning (DL) QSAR model that combines a recurrent neural network with regression model. By utilizing the ability of the regression model to determine quantitative relationships between peptide sequences and biofilm inhibition and combining this with the sequence-based RNN in antimicrobial classification, this approach effectively predicted activity with high training accuracy to the level of 90% in 3-fold cross-validation. This robust predictive capability resulted in the identification of novel short cationic peptides with substantially improved antibiofilm and antimicrobial properties.

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.19.644165; this version posted March 19, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The peptides found by the constructed approach demonstrated more effective biofilm inhibition and antimicrobial activity compared to the training peptides, with stronger inhibition ability on average in antibiofilm and antimicrobial assays [Figure 3]. The ability of the DL methodology to efficiently and simultaneously screen large peptide datasets for multiple therapeutic endpoints can enhance the precision and effectiveness of AMP discovery.

In conclusion, this work highlights the potential of AI-based approaches to accelerate the identification of next-generation therapeutics discovery, paving the way for novel strategies to combat resistant bacterial infections.

References

1. Lei, Jun, et al. "The antimicrobial peptides and their potential clinical applications." American Journal of Translational Research 11.7 (2019): 3919.
2. Li, Jianguo, et al. "Membrane active antimicrobial peptides: translating mechanistic insights to design." Frontiers in Neuroscience 11 (2017): 73.
3. Nedyalkova, Miroslava, et al. "Progress and future of the computational design of antimicrobial peptides (AMPs): bio-inspired functional molecules." Digital Discovery 3.1 (2024): 9-22.
4. Cherkasov, Artem. "Inductive QSAR descriptors. Distinguishing compounds with antibacterial activity by artificial neural networks." International Journal of Molecular Sciences 6.1 (2005): 63-86.
5. Tropsha, Alexander, et al. "Integrating QSAR modelling and deep learning in drug discovery: the emergence of deep QSAR." Nature Reviews Drug Discovery 23.2 (2024): 141-155.
6. Mansour, Sarah C., César de la Fuente□Núñez, and Robert EW Hancock. "Peptide IDR□1018: modulating the immune system and targeting bacterial biofilms to treat antibiotic□resistant bacterial infections." Journal of Peptide Science 21.5 (2015): 323-329.
7. Barathe, Pramod, et al. "Antibiotic pollution and associated antimicrobial resistance in the environment." Journal of Hazardous Materials Letters (2024): 100105.
8. de Ilurdoz, M. Sagaseta, J. Jaime Sadhwani, and J. Vaswani Reboso. "Antibiotic removal processes from water & wastewater for the protection of the aquatic environment-a review." Journal of Water

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.19.644165; this version posted March 19, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Process Engineering 45 (2022): 102474.

9. Zheng, Yong Fang, et al. "de Novo-designed antimicrobial peptides with broad-spectrum antimicrobial potency and rapid wound disinfection." Colloids and Surfaces A: Physicochemical and Engineering Aspects 667 (2023): 131355.
10. Zhang, Qi-Yu, et al. "Antimicrobial peptides: mechanism of action, activity and clinical potential." Military Medical Research 8 (2021): 1-25.
11. Nissan, Nour, et al. "Future perspective: harnessing the power of artificial intelligence in the generation of new peptide drugs." Biomolecules 14.10 (2024): 1303.
12. Swan, Anna Louise, et al. "Application of machine learning to proteomics data: classification and biomarker identification in postgenomics biology." Omics 17.12 (2013): 595-610.
13. Dockès, Jérôme, Gaël Varoquaux, and Jean-Baptiste Poline. "Preventing dataset shift from breaking machine-learning biomarkers." GigaScience 10.9 (2021): giab055.
14. Xie, Ying, et al. "Early lung cancer diagnostic biomarker discovery by machine learning methods." Translational Oncology 14.1 (2021): 100907.
15. You, Jiaying, Robert D. McLeod, and Pingzhao Hu. "Predicting drug-target interaction network using deep learning model." Computational Biology and Chemistry 80 (2019): 90-101.
16. Echle, Amelie, et al. "Deep learning in cancer pathology: a new generation of clinical biomarkers." British Journal of Cancer 124.4 (2021): 686-696.
17. Cui, Lei, et al. "A deep learning-based framework for lung cancer survival analysis with biomarker interpretation." BMC Bioinformatics 21 (2020): 1-14.
18. Alpaydin, Ethem. Machine learning. MIT press, 2021.
19. Hunt, Earl B. Artificial intelligence. Academic Press, 2014.
20. Cunningham, Pádraig, Matthieu Cord, and Sarah Jane Delany. "Supervised learning." Machine learning techniques for multimedia: case studies on organization and retrieval. Berlin, Heidelberg: Springer Berlin Heidelberg, 2008. 21-49.
21. Hastie, Trevor, et al. "Unsupervised learning." In: The elements of statistical learning: Data mining, inference, and prediction (2009): 485-58
22. Carcillo, Fabrizio, et al. "Combining unsupervised and supervised learning in credit card fraud detection." Information Sciences 557 (2021): 317-331.
23. Yu, Junliang, et al. "Self-supervised learning for recommender systems: A survey." IEEE Transactions on Knowledge and Data

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.19.644165; this version posted March 19, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Engineering 36.1 (2023): 335-355.

24. Marcos-Zambrano, Laura Judith, et al. "Applications of machine learning in human microbiome studies: a review on feature selection, biomarker identification, disease prediction and treatment." Frontiers in Microbiology 12 (2021): 634511.
25. Wang, Yunpeng, et al. "Pre-training with asynchronous supervised learning for reinforcement learning based autonomous driving." Frontiers of Information Technology & Electronic Engineering 22.5 (2021): 673-686.
26. Yoshida, Mari, et al. "Using evolutionary algorithms and machine learning to explore sequence space for the discovery of antimicrobial peptides." Chem 4.3 (2018): 533-543.
27. Sarker, Iqbal H. "Deep learning: a comprehensive overview on techniques, taxonomy, applications and research directions." SN Computer Science 2.6 (2021): 1-20.
28. Sherstinsky, Alex. "Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network." Physica D: Nonlinear Phenomena 404 (2020): 132306.
29. Wang, Fei, et al. "Phosphorylation-dependent positive feedback on the oxytocin receptor through the kinase PKD1 contributes to long-term social memory." Science Signaling 15.719 (2022): eabd0033.
30. Huang, Guohua, et al. "LSTMCNNsucc: A bidirectional LSTM and CNN□based deep learning method for predicting lysine succinylation sites." BioMed Research International 2021.1 (2021): 9923112.
31. Chung, Cheuk To, et al. "Machine learning techniques for arrhythmic risk stratification: a review of the literature." International Journal of Arrhythmia 23.1 (2022): 10.
32. Hochreiter, Sepp. "Recurrent neural net learning and vanishing gradient." International Journal of Uncertainity, Fuzziness and Knowledge-Based Systems 6.2 (1998): 107-116.
33. Li, You, et al. "MPMABP: a CNN and Bi-LSTM-Based method for predicting multi-activities of bioactive peptides." Pharmaceuticals 15.6 (2022): 707.
34. UniProt Consortium. "UniProt: a hub for protein information." Nucleic Acids Research 43.D1 (2015): D204-D212.
35. Willmott, Cort J., and Kenji Matsuura. "Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance." Climate Research 30.1 (2005): 79-82.
36. Etayash, Hashem, Evan F. Haney, and Robert EW Hancock. "Assessing biofilm inhibition and immunomodulatory activity of small amounts of synthetic host defense peptides synthesized using

bioRxiv preprint doi: https://doi.org/10.1101/2025.03.19.644165; this version posted March 19, 2025. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

SPOT-array technology." Nature Protocols 16.4 (2021): 1850-1870.

37. Shi, Guobang, et al. "DRAMP 3.0: an enhanced comprehensive data repository of antimicrobial peptides." Nucleic Acids Research 50.D1 (2022): D488-D496.
38. Lin, Tzu-Tang, et al. "AI4AMP: an antimicrobial peptide predictor using physicochemical property-based encoding method and deep learning." Msystems 6.6 (2021): e00299-21.
39. Pirtskhalava, Malak, et al. "DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics." Nucleic Acids Research 49.D1 (2021): D288-D297.
40. Al-Shehari, Taher, and Rakan A. Alsowail. "An insider data leakage detection using one-hot encoding, synthetic minority oversampling and machine learning techniques." Entropy 23.10 (2021): 1258.
41. Hancock, R.E.W., M. Alford, and E.F. Haney. “Antibiofilm activity of host defence peptides: Complexity provides opportunities.” Nature Microbiology Reviews 19 (2021):786-797.