Comprehensive Machine Learning Analysis of Long and Middle Peptides: Supervised and Unsupervised Approaches

Ahmed El-Gabry
Deptment of Healthcare Engineering
Fuclty of Enginering
Cairo University
Giza , Egypt

Antonious Atef Saleh
Deptment of Healthcare Engineering
Fuclty of Enginering
Cairo University
Giza , Egypt

Omar El Saeed
Deptment of Healthcare Engineering
Fuclty of Enginering
Cairo University
Giza , Egypt

***Abstract*—**

This study investigates antimicrobial peptides (AMPs), pivotal in combating infections, using accessible machine learning methods. We examined long, medium, and short peptides, focusing on specific features. Initially, supervised classification, guided by a reference paper from fellow researchers in our department, was employed to analyze peptides across several features. This approach provided insights into the effectiveness of these peptides. Subsequently, we adopted unsupervised learning techniques, utilizing tools such as SVM (Support Vector Machines), RF (Random Forest), and KNN (K-Nearest Neighbors). Our findings unveil new insights into the peptides, revealing both anticipated and unexpected patterns. While the supervised approach helped us understand the known characteristics, unsupervised learning allowed for the discovery of hidden analogies and patterns not considered by traditional chemical analysis. This work is significant as it deepens our comprehension of AMPs, paving the way for improved treatments for infections. The study underscores the synergy between machine learning and biochemical insights in the exploration of peptide functionality.

***Keywords*—**

Antimicrobial Peptides (AMPs), Machine Learning Techniques, Biological Insights, Biochemical analysis, Therapeutic peptides, Supervised vs. unsupervised learning, iLearn Plus tool, Protein sequence analysis, Biomedical Research

I. INTRODUCTION

Antimicrobial peptides (AMPs) are crucial in our defense against infectious diseases, with their diverse and potent threptic properties. This study discovers computational techniques and machine learning algorithms, including Support Vector Machines (SVM), Random Forest (RF), k-Nearest Neighbors (KNN), alongside deep learning models like Convolutional Neural Networks (CNN) and Artificial Neural Networks (ANN), to predict AMP efficacy. We explore AMPs of various lengths - long, middle, and short - utilizing iLearn Plus for feature extraction, including Amino Acid Composition (AAC), Grouped Amino Acid Composition (GAAC), Geary Auto Correlation (GAC) and Composition, Transition, and Distribution (CTD), all through analyzing features along the paths of Amino Acid Composition (AAC): This feature type represents the frequency of each amino acid in a peptide sequence. It gives a basic overview of the peptide's composition. Dipeptide Composition: This involves the frequency of pairs of amino acids (dipeptides) in the peptide sequence. It provides information on the local sequence patterns. Physicochemical Properties: These features include various physical and chemical properties of amino acids, such as hydrophobicity, charge, molecular weight, etc. They can help in understanding how peptides interact with microbial membranes or immune cells. Molecular Descriptors: These include structural features of peptides, like molecular weight, polarity, or shape descriptors. They are more common in cheminformatics approaches. Sequence-Based Features: These can include motifs, patterns, or specific sequence alignments that are known to be important in the function of AMPs. Secondary Structure: The structural features related to the shape of the peptide, like alpha-helix or beta-sheet content. N- and C-Terminal Features: Properties or compositions specific to the N-terminal (beginning) and C-terminal (end) regions of the peptide, which can be crucial for their function.

Our approach extends beyond conventional methods by integrating both supervised and unsupervised learning. The supervised models focus on classifying AMPs based on their antimicrobial activities, enhanced by feature selection and scaling techniques. In parallel, we delve into the realms of unsupervised learning with DBSCAN and K-means++ algorithms, aiming to discover hidden patterns and clusters in the peptide data. This dual approach allows for a more profound understanding of AMP properties and their potential therapeutic roles.

Building upon the existing frameworks, we have made significant progress in refining the predictive models and enhancing their accuracy. Our work has led to the identification of novel peptide sequences with promising antimicrobial traits, contributing valuable insights to the field of antimicrobial research and offering new directions in the fight against antibiotic resistance.

II. RELATED WORK

*A. Main paper contribution *Anti-Microbial activity prediction**

The study 'Machine Learning Prediction of Antimicrobial Peptides' provides a significant contribution to the field of AMP prediction using machine learning. The authors effectively utilize a range of algorithms, including SVM, RF, and KNN, and incorporate deep learning models like CNNs and ANNs. Their use of a diverse dataset covering different peptide lengths and the application of iLearn Plus for feature extraction is commendable. The integration of supervised and

unsupervised learning techniques offers a comprehensive approach to understanding AMPs.

However, the study has limitations that our research aims to address. One notable aspect is the focus on a specific set of peptide descriptors, which may not encompass the full spectrum of AMP characteristics. Additionally, while the unsupervised learning approach offers insights, there is room for further exploration in this area, particularly in identifying novel peptide sequences. Our work builds upon these foundations, aiming to expand the understanding of AMP properties and enhance the predictive accuracy of these models.

B. Further related work on the matter

1. **Deep Learning for Drug Discovery:** This study leverages deep learning and molecular dynamics to efficiently develop antimicrobial compounds, showing promising results in combating antibiotic resistance.4. Machine Learning in Peptide Activity: This review covers recent machine learning approaches for peptide activity prediction, highlighting the importance of method selection based on dataset and research goals.

2. Multi-Functional Peptide Prediction: Utilizing deep learning, this research improves the prediction of therapeutic peptides, though it lacks comparative analysis with existing methods.

3. Challenges in Biosciences via Deep Learning: The paper discusses deep learning applications in biosciences, highlighting significant advancements and ongoing challenges in protein structure prediction and genome engineering.

4. Machine Learning in Peptide Activity: This review covers recent machine learning approaches for peptide activity prediction, highlighting the importance of method selection based on dataset and research goals.

5. Language Models in AMP Classification: Introducing a novel approach using language models and CNNs for antimicrobial peptide classification, this study demonstrates high accuracy but lacks in-depth limitations analysis.

6. Antibiotic Discovery using Deep Learning: The research presents a deep learning model for new antimicrobial activities, emphasizing the need for further properties investigation of the discovered peptides.

7. Antimicrobial Peptide Recognition via Genetic Programming: Focusing on gram-positive and gram-negative AMPs, this paper combines evolutionary algorithms with feature construction, noting the computational intensity of the method.

8. DNN Classifier in AMP Recognition: This study introduces a deep neural network classifier for AMP recognition, showcasing effectiveness but requiring extensive computational resources.

9. AMP Prediction with Machine Learning: Highlighting the importance of machine learning for AMP prediction, this paper discusses the role of AMP databases in enhancing peptide discovery.

10. Deep Learning in AMP Discovery: Reviewing the use of deep learning for AMP prediction and design, this work covers the latest advancements and challenges, including data scarcity and the need for explainable AI.

11. Machine Learning and AMPs: This review assesses machine learning's capability in identifying key features of antimicrobial peptides (AMPs) and their potential for designing new AMPs, acknowledging its limitations in differentiating between antimicrobial and membrane activities.

12. Machine Learning for Membrane-Active Peptides: The paper discusses using machine learning, particularly support vector machines, to discover and design membrane-active peptides, highlighting the need for experimental validation.

13. ML-Assisted Peptide Design for Drug Discovery: Focusing on drug discovery, this study leverages machine learning to identify peptides with high biological activity, showcasing an efficient algorithm for sorting peptides.

14. AMP Databases and Computational Tools: This review highlights the role of computational methods and databases in the prediction and design of new AMPs, emphasizing the need for improved machine learning algorithms.

15. MAMPs-Pred for Antimicrobial Peptides: The study presents a machine learning approach called MAMPs-Pred to identify AMPs and their functions, demonstrating improved accuracy over existing methods.

16. Deep Learning in AMP Recognition: Researchers develop a deep learning model incorporating convolutional and recurrent layers to recognize AMPs, addressing the challenge of bacterial resistance to antibiotics.

17. Deep Learning for AMP Design: This paper details the use of deep learning, particularly LSTM models, for designing short AMPs targeting Gram-negative bacteria, highlighting the challenges in AMP development.

18. ML Techniques for AMP Prediction: The article discusses the growing need for computational models to predict AMPs, given their diverse biological activities and mechanisms.

19. Computational Approaches for AMP Discovery: This article explores AI and ML algorithms in discovering and designing AMPs, suggesting the need for non-standard feature exploration and big data platforms.

20. Multi-Scale Convolutional Network for AMP Identification: This study introduces a deep learning model with a multi-scale convolutional network for identifying sequences, demonstrating superior accuracy over traditional models.

III. SYSTEM WIDE ARCHETECTURE

The figure illustrates a comprehensive block diagram outlining the software architecture designed for a protein classification system, with our recent enhancements incorporating advanced unsupervised machine learning models. The system unfolds across three primary layers: the data layer, the processing layer, and the output layer three core layers drive the functionality: the data layer, the processing layer, and the output layer.

Data Layer:

Responsible for the storage and retrieval of protein sequences, the data layer employs a database structure. Specifically, protein sequences, including those pertinent to antimicrobial peptides (AMPs) from our recent project, find residence in this database for efficient access by the processing layer.

Processing Layer:

This layer is tasked with the classification of protein sequences. Implemented are three distinct models, each rigorously trained on varied datasets. These models collectively contribute to predicting the class of novel protein sequences. In our recent project integration, specialized models were incorporated to enhance the system's predictive capabilities, particularly in AMP classification.

Output Layer:

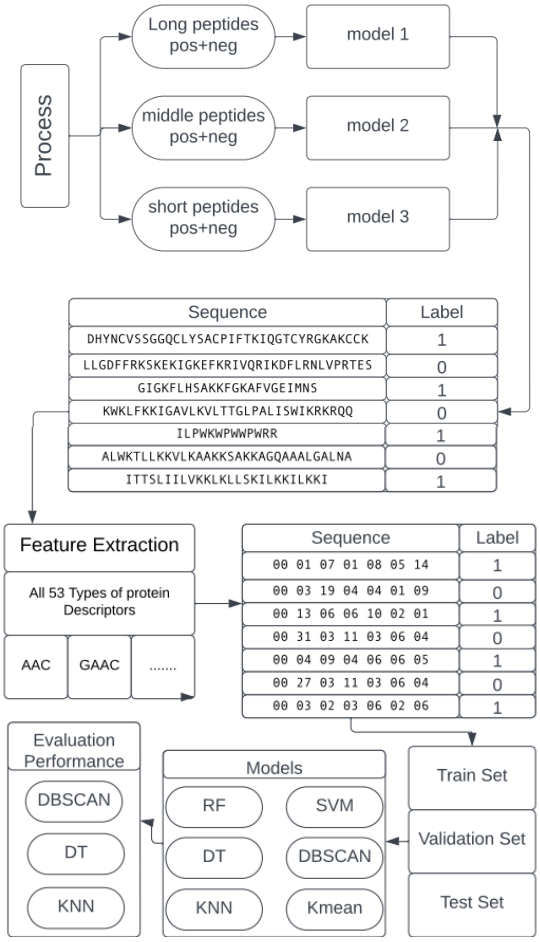

Fig.1 Proposed framework

The output layer plays a pivotal role in presenting classification results. The outcomes, displayed in text format, can be seamlessly exported to a file for further analysis or

Several advantages characterize this architecture:

Scalability: The system can adeptly handle an increasing volume of protein sequences and models.

Flexibility: Beyond its current application, the architecture is adaptable for various protein classification tasks.

Accuracy: The utilization of three models, trained on distinct datasets, enhances the precision of classification, a feature augmented by our recent project focus on AMPs.

However, challenges persist in Complexity: The inherent complexity of the architecture may pose challenges in comprehension and maintenance. Data Requirements: A substantial dataset of protein sequences is essential for optimal system performance. Training Time: Training the models may demand a significant time investment. Integration into external systems..

IV. METHODOLOGY:

A. Data sets

In this study, we utilized a comprehensive dataset identical to that used in our reference paper. This dataset is meticulously curated from a well-established source, encompassing a broad spectrum of antimicrobial peptides (AMPs) categorized by their amino acid length. The dataset includes three distinct classes: long peptides (50-100 residues), middle peptides (10-50 residues), and short peptides (less than 10 residues). Each category comprises both positive and negative samples, reflecting peptides with and without antimicrobial properties, respectively. The data for long peptides includes 3,776 positive and 29,640 negative samples, middle peptides consist of 17,144 positive and 41,713 negative samples, while short peptides are represented by 5,695 positive and 2,737 negative samples. This distribution was crucial in ensuring a balanced approach to machine learning model training, thus enhancing the robustness and reliability of our predictions. Employing this dataset allowed us to mirror the comprehensive analysis conducted in the reference paper, providing a solid foundation for comparative study and further exploration in the realm of AMPs.

B. Feature Scaling

Following peptide sequence extraction, converting sequences into numerical descriptors, including amino acid content, physiochemical properties, hydrophobicity, and net charge, we applied feature scaling. This crucial preprocessing step standardizes features, ensuring a consistent scale. In our study, we employed z-score normalization, setting each feature's mean to 0 and standard deviation to 1. This process enhances machine learning model efficiency and accuracy by placing all features on a common scale. Standardization prevents larger-value features from dominating the training process, promoting fair contribution from all descriptors in predicting antimicrobial activity. Incorporating feature scaling optimizes model performance, improving accuracy and robustness in predicting relationships between peptide descriptors and antimicrobial activity..

C. Feature selection

To address the high dimensionality challenges of our dataset, characterized by numerous descriptors, we applied feature selection techniques to pinpoint the most informative and relevant features for our model.

One such technique employed was Random Forest feature selection, an ensemble machine learning algorithm. This method constructs multiple decision trees and gauges the importance of each feature by assessing its impact on model accuracy when removed. Ranking features based on their contribution to predictive performance helped identify the most significant ones, minimizing the risk of overfitting or underfitting.

We also utilized Chi-squared feature selection, a statistical test measuring dependence between two categorical variables. This technique evaluated the relevance of each feature to the class label, aiding in the selection of informative descriptors closely linked to the target variable.

Incorporating these feature selection methods streamlined the feature set, retaining only the most critical features. This not only enhanced model performance by reducing noise but also improved interpretability, offering insights into specific aspects of peptide sequences significantly contributing to their antimicrobial activity.

D. Equations

1. **Accuracy (ACC):** $\frac{TP+TN}{TP+TN+FP+FN}$

TP=True Positives TN=True Negatives

FP=False Positives FN=False Negatives

2. **Recall (Sensitivity, True Positive Rate, TPR):** $\frac{TP}{TP+FN}$

3. **Precision (Positive Predictive Value, PPV):** $\frac{TP}{TP+FT}$

4. **Receiver Operating Characteristic (ROC) Curve and Area Under the Curve (AUC):**

The ROC curve is a plot of TPR (Recall) against FPR (1 - Specificity) at different classification thresholds.

AUC = Area under the ROC curve

AUC represents the overall performance of the model in distinguishing between classes.

For unsupervised learning metrics:

1. **Inertia:** Sum of squared distances of each data point in a cluster to its cluster's centroid

2. **Silhouette Score:** $\frac{b-a}{max(a,b)}$

a = average distance between a data point and all other points in the same cluster

b = average distance between a points and all points in the nearest neighboring cluster

Higher Silhouette Scores indicate better-defined clusters, with values ranging from -1 to 1. A score close to 1 indicates well-clustered data points.

Specificity $= \frac{TN}{(TN+FP)}$

The Area Under the Precision-Recall Curve (AUC-PR)

A higher AUC-PR indicates better performance, especially in situations where the class distribution is imbalanced.

V. RESULTS

Supervised Learning Models - Long Peptides:

The ML models on the Long Peptides dataset demonstrated the following top performers:

SVM with rbf Kernel:

Accuracy	**0.85**
Precision (Positive)	0.82
Recall (Positive)	0.85

Random Forest:

Accuracy	**0.84**
Precision (Positive)	not provided
Recall (Positive)	not provided

Ada Boost:

Accuracy	**0.83**
Precision (Positive)	not provided
Recall (Positive)	not provided

KNN with k=5:

Accuracy	**0.82**
Precision (Positive)	not provided
Recall (Positive)	not provided

Decision Tree:

Accuracy	**0.77**
Precision (Positive)	not provided
Recall (Positive)	not provided

Remarks:
The SVM with rbf kernel model demonstrated the highest accuracy, indicating its suitability for the task.
Both the Random Forest and Ada Boost models showed high accuracies, suggesting their effectiveness in predicting antimicrobial activity.
The Decision Tree model had the lowest accuracy, implying it may not be as effective for this task.
Precision and recall scores for each model indicate good performance in identifying both positive and negative

examples.
Overall, the models, especially SVM with rbf kernel, Random Forest, Ada Boost, and KNN with k=5, exhibit promise for predicting antimicrobial activity, offering potential applications in antibiotic development.

Unsupervised Learning Models - Long Peptides:

Table 1: Decision Tree Classifier:

Accuracy			**77%**
Precision	Class 0: 74%		Class 1: 81%
Recall	Class 0: 82%		Class 1: 71%
F1-Score:	Class 0: 78%		Class 1: 76%
Confusion Matrix:			
TP: 614	FP: 131	FN: 219	TN: 547

Table 2: Random Forest Classifier:

Accuracy		**83%**	
Precision	Class 0: 81%		Class 1: 85%
Recall	Class 0: 86%		Class 1: 81%
F1-Score:	Class 0: 83%		Class 1: 83%
	Confusion Matrix:		
TP: 639	FP: 106	FN: 147	TN: 619

Table 3: Naïve Bayes Classifier:

Accuracy		**70%**	
Precision	Class 0: 65%		Class 1: 82%
Recall	Class 0: 89%		Class 1: 51%
F1-Score:	Class 0: 75%		Class 1: 63%
	Confusion Matrix:		
TP: 680	FP: 83	FN: 364	TN: 384

Table 3: K-Nearest Neighbors (KNN) Classifier

Accuracy		**82.26%**	
Precision	Class 0: 82%		Class 1: 83%
Recall	Class 0: 85%		Class 1: 79%
F1-Score:	Class 0: 83%		Class 1: 81%
	Confusion Matrix:		
TP: 671	FP: 119	FN: 149	TN: 572

Table 4: Support Vector Machine (SVM) Classifier:

Accuracy		**85.31%**	
Precision	Class 0: 85.86%		Class 1: 84.70%
Recall	Class 0: 86.08%		Class 1: 84.47%
F1-Score:	Class 0: 83%		Class 1: 81%
	Confusion Matrix:		
TP: 680	FP: 110	FN: 112	TN: 609

Table 5: ROC Analysis:

KNN: AUC	**0.89**
SVM: AUC	0.91
AdaBoost: AUC	0.89
Decision Tree: AUC	0.84
Random Forest: AUC	0.91
Naive Bayes: AUC	0.80

Table 6: Long Peptides results

	Accuracy	**Precis (+)**	**Recall (+)**	**Precis (-)**	**Recall (-)**
KNN	82.2%	83%	79%	82%	85%
SVM	85.3%	84.7%	84.4%	85.8%	86%
AdaBoost	82%	82%	80%	82%	84%
Random Forest	83%	85%	81%	81%	86%
Decision Tree	77%	81%	71%	74%	82%

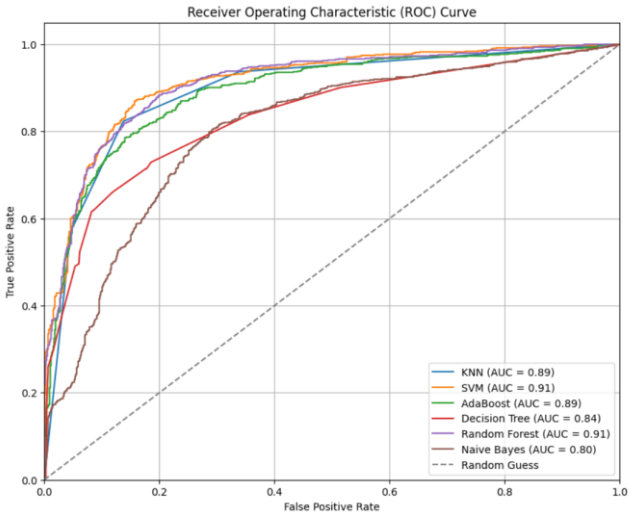

Fig. 2 ROC graph of long peptides

	Accuracy	Precis (+)	Recall (+)	Precis (-)	Recall (-)
Random Forest	70%	70%	71%	70%	69%
SVM	67.9%	67.7%	67%	68%	68.7%
Ada Boost	67%	66%	70%	68%	63%

Table 2: Middle Peptides results

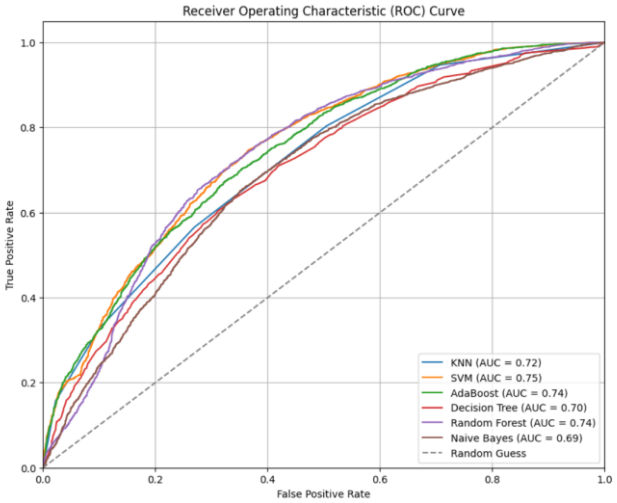

Fig. 3 ROC graph for middle peptides

	Accuracy	Precis (+)	Recall (+)	Precis (-)	Recall (-)
Random Forest	70.6%	75%	63%	67%	79%
SVM	73.4%	74.3%	72.4%	72.4%	73.9%
Ada Boost	72%	74%	71%	71%	74%

Table 3: Short Peptides results

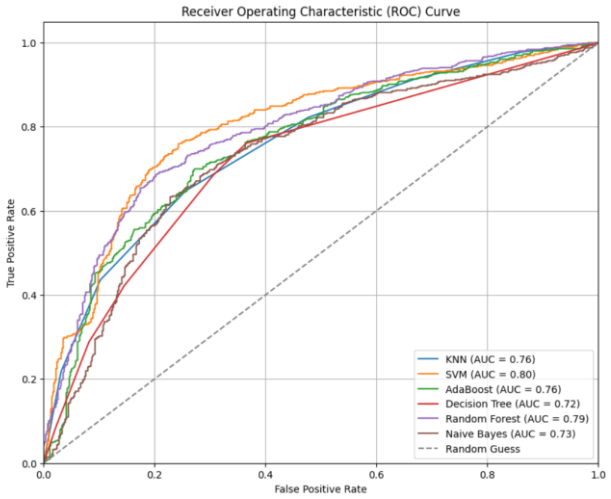

Fig.4 Short peptide graph

VI. Conclustion

The SVM model is more sensitive than the other models, which means that it is better at identifying the antimicrobial activity of the peptides. However, it is also more prone to false positives. • The KNN model is less sensitive than the SVM model, but it is also less prone to false positives. This makes it a good choice for applications where false positives are costly. • The Decision Tree model is the least sensitive of the models shown, but it is also the least prone to false positives. This makes it a good choice for applications where it is important.

References

[1] Géron, A. (2017). Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems. http://cds.cern.ch/record/2699693

[2] Kelleher, J. D., Mac Namee, B., & Darcy, A. B. (2015). Fundamentals of Machine learning for Predictive data analytics: algorithms, worked examples, and case studies. https://dl.acm.org/citation.cfm?id=2815672

[3] Alpaydin, E. (2016). Machine Learning : the new AI. https://international.scholarvox.com/book/88841730

[4] iLearnPlus Web. (n.d.). Ilearnplus.erc.monash.edu. Retrieved August 2, 2023, from https://ilearnplus.erc.monash.edu/

[5] Das, P., Sercu, T., Wadhawan, K., Padhi, I., Gehrmann, S., Cipcigan, F., Chenthamarakshan, V., Strobelt, H., Santos, C. D., Chen, P., Yang, Y. Y., Tan, J. P. K., Hedrick, J. L., Crain, J., & Mojsilovic, A. (2021). Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. Nature Biomedical Engineering, 5(6), 613– 623. https://doi.org/10.1038/s41551-021-00689-x

[6] Fan, H., Yan, W., Wang, L., Liu, J., Bin, Y., & Xia, J. (2023). Deep learning-based multifunctional therapeutic peptides prediction with a multi-label focal dice loss function. Bioinformatics, 39(6). https://doi.org/10.1093/bioinformatics/btad334

[7] Sapoval, N., Aghazadeh, A., Nute, M., Antunes, D. A., Balaji, A., Baraniuk, R. G., Barberan, C., Dannenfelser, R., Dun, C., Edrisi, M., Elworth, R. a. L., Kille, B., Kyrillidis, A., Nakhleh, L., Wolfe, C. R., Yan, Z., Yao, V., & Treangen, T. J. (2022). Current progress and open challenges for applying deep learning across the biosciences. Nature Communications, 13(1). https://doi.org/10.1038/s41467-022-29268-7

[8] Wu, Q., Ke, H., Dong-Li, L., Wang, Q., Fang, J., & Zhou, J. (2019). Recent progress in machine learning-based prediction of peptide activity for drug discovery. Current Topics in Medicinal Chemistry, 19(1), 4–16. https://doi.org/10.2174/1568026619666190122151634

[9] Dee, W. (2022). LMPred: predicting antimicrobial peptides using pre-trained language models and deep learning. Bioinformatics Advances, 2(1). https://doi.org/10.1093/bioadv/vbac021

[10] Singh, V., Shrivastava, S., Singh, S. K., Kumar, A., & Saxena, S. (2023). Multi-scale temporal convolutional networks and continual learning based in silico discovery of alternative antibiotics to combat multi-drug resistance. Expert Systems With Applications, 215, 119295. https://doi.org/10.1016/j.eswa.2022.119295

[11] Veltri, D., Kamath, U., & Shehu, A. (2017). Improving Recognition of Antimicrobial Peptides and Target Selectivity through Machine Learning and Genetic Programming. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 14(2), 300–313. https://doi.org/10.1109/tcbb.2015.2462364

[12] Veltri, D., Kamath, U., & Shehu, A. (2018). Deep learning improves antimicrobial peptide recognition. Bioinformatics, 34(16), 2740–2747. https://doi.org/10.1093/bioinformatics/bty179

[13] Wang, G., Vaisman, I. I., & Van Hoek, M. L. (2022). Machine learning prediction of antimicrobial peptides. In Springer eBooks (pp. 1–37). https://doi.org/10.1007/978-1-0716-1855-4_1

[14] Yan, J., Cai, J., Zhang, B., Wang, Y., Wong, D. F., & Siu, S. W. I. (2022). Recent progress in the discovery and design of antimicrobial peptides using traditional machine learning and deep learning. Antibiotics, 11(10), 1451. https://doi.org/10.3390/antibiotics11101451

[15] Lee, E. Y., Lee, M. W., Fulan, B. M., Ferguson, A. L., & Wong, G. C. L. (2017). What can machine learning do for antimicrobial peptides, and what can antimicrobial peptides do for machine learning? Interface Focus, 7(6), 20160153. https://doi.org/10.1098/rsfs.2016.0153

[16] Lee, E. Y., Wong, G. C. L., & Ferguson, A. L. (2018). Machine learning-enabled discovery and design of membrane-active peptides. Bioorganic & Medicinal Chemistry, 26(10), 2708–2718. https://doi.org/10.1016/j.bmc.2017.07.012

[17] Giguère, S., Laviolette, F., Marchand, M., Tremblay, D. M., Moineau, S., Liang, X., Biron, E., & Corbeil, J. (2015). Machine learning assisted design of highly active peptides for

drug discovery. PLOS Computational Biology, 11(4), e1004074. https://doi.org/10.1371/journal.pcbi.1004074

[18] Abdolmaleki, P., Mohammadi, N., Allahverdi, A., Khalili, E., & Abdolmaleki, P. (2022). A review on antimicrobial peptides databases and the computational tools. Database, 2022. https://doi.org/10.1093/database/baac011

[19] Lin, Y., Cai, Y., Liu, J., Lin, C., & Liu, X. (2019). An advanced approach to identify antimicrobial peptides and their function types for penaeus through machine learning strategies. BMC Bioinformatics, 20(S8). https://doi.org/10.1186/s12859-019-2766-9

[20] Veltri, D., Kamath, U., & Shehu, A. (2018b). Deep learning improves antimicrobial peptide recognition. Bioinformatics, 34(16), 2740–2747. https://doi.org/10.1093/bioinformatics/bty179

https://doi.org/10.3390/biom11030471