

Prediction Of Antimicrobial Peptides Using Machine Learning Approaches

Project report submitted in partial fulfilment of the requirement for the
degree of

Bachelor of Technology

in

Biotechnology

Submitted by

Drishti Awasthi (201817)

Under the supervision of

Dr. Jitendraa Vashistt



Department of Biotechnology and Bioinformatics

**Jaypee University of Information Technology Waknaghat, Solan-
173234, Himachal Pradesh**



Certificate

This is to certify that the project entitled “Prediction of Antimicrobial Peptides using Machine learning Approaches”, submitted by Drishti Awasthi in partial fulfilment of the award of the degree Bachelor of Technology in Biotechnology to Jaypee University of Information Technology, Waknaghat, Solan has been made under my supervision. This work has not been submitted partially or wholly to any other university or institute for the award of any other degree, diploma or such other titles.

Dr. Jitendraa Vashistt (Guide)

Associate Professor

Candidate Declaration

I hereby declare that the work presented in this report entitled '**Prediction of Antimicrobial Peptides using Machine Learning Approaches**' in partial fulfilment of the requirements for the award of the degree of **Bachelor of Technology in Biotechnology** submitted in the Department of Biotechnology and Bioinformatics, Jaypee University of Information Technology, Waknaghat is an authentic record of my own work carried out over a period from August 2023 to May 2024 under the supervision of **Dr. Jitendraa Vashistt** (Associate Professor, Department of Biotechnology and Bioinformatics). The matter embodied in the report has not been submitted for the award of any other degree or diploma.

(Student Signature)

Drishti Awasthi

201817

This is to certify that the above statement made by the candidate is true to the best of my knowledge.

(Supervisor Signature)

Dr. Jitendraa Vashistt,

Associate Professor,

Department of Biotechnology and Bioinformatics

Dated:

Acknowledgement

I would like to express my heartiest gratefulness to God for his divine blessing to make it possible to complete the project work successfully. I am grateful and wish my profound indebtedness to Dr Jitendraa Vashistt, Associate Professor, Department of BT& BI, Jaypee University of Information Technology, Waknaghat. The deep knowledge & keen interest of my supervisor in the field of “Molecular biology and Biochemistry” gave the idea and passion to integrate technology with biology. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, and reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

I extend my gratitude to Professor Sudhir Kumar, Head of Department BT & BI, at Jaypee University of Information Technology for his constant encouragement and motivation that enthused me to learn and process knowledge with zest and determination. I would like to express our sincere appreciation to Professor Dr. Shruti Jain, Professor and Associate Dean(Innovation) at the Jaypee University of Information Technology, for her insightful guidance regarding Machine Learning approaches that were implemented during the course of my project. I would also like to express my gratitude to Dr. Shikha Mittal, Assistant professor, at Jaypee University of Information Technology for helping me out with all the bioinformatics and feature extraction processes. I would also generously acknowledge each one of those individuals who have helped me straightforwardly or in a roundabout way in making this project a win. In this unique situation, I might want to thank the various staff individuals, both educating and non-instructing, which have developed their convenient help and facilitated my undertaking. No expression of appreciation is complete without recognition of the prayers, good wishes, advice and moral support of my affectionate parents which helped us immensely to achieve my goal.

Name: Drishti Awasthi

Roll Number: 201817

Table of Content

Title	Page No.
Certificate	II
Candidate Declaration	III
Acknowledgement	IV
Table Of Content	V
List of Figures	VI
List of Tables	VII
List of Abbreviations	VIII
Abstract	IX
Chapter 1: Introduction	10
Chapter 2: Literature Review	14
Chapter 3: Materials and Methods	23
Chapter 4: Results	31
Chapter 5: Discussions	40
Chapter 6: Conclusion	42
Plagiarism Certificate	43
References	44

List of Figures

Figure No.	Caption
1	Flowchart of model implementation
2	Pie chart on AMP distribution
3	Summary statistics of dataset
4	Python code for hydrophobicity calculator
5	Scrapping of data using Selenium
6	Hydrophobicity of each amino acid
7	Net charge calculating code
8	Classification report of logistic regression
9	Prediction of AMP/ non-AMP
10	Antibacterial peptides predicted by ML model
11	Antigram negative peptide prediction
12	Gram variable peptide prediction
13	Antigram positive peptide prediction
14	Random sequence prediction obtained from different databases
15	Accuracy comparison voting of ML models
16	Voting example for a sequence
17	Accuracy score of ML models
18	Prediction made for each sequence

List of Tables

Table No.	Caption
1	Hydrophobic amino acids and their value
2	Crucial physico-chemical properties.

List of Abbreviations

1. AMP: Antimicrobial Peptide
2. ABP: Antibacterial Peptides
3. ML-KNN: Multi label K-nearest neighbour
4. LSTM: Long short term Memory
5. DT: Decision Trees
6. RF: Random Forest
7. CNN: Convolutional Neural Network
8. ABP3Finder: Antibacterial peptide 3 Finder
9. AGP: Antigram positive
10. AGN: Antigram negative

Abstract

Anti Microbial research has always been an interesting field in which gram-positive and gram-negative bacteria are the primary targets of the majority of the currently available techniques used for predicting antimicrobial peptides (AMPs). The approach we present in this work enables us to forecast AMPs against gram-positive, gram-negative, as well as bacteria in general. In this project the APD3 (Animicrobial Peptide Database), PDB (Protein Databank), UniProt (Universal Protein Resource) databases were used to curate a dataset for small peptides and their physico-chemical properties were calculated using hydrophobic and net charge scales at an adequate pH. The project achieves 97.07% accuracy for predicting if a peptide is an AMP or not and 88% accuracy for predicting antigram-positive, antigram-negative or antigram-variable peptides. The exceptional part of this implementation multi label classification using TF-IDF (Term Frequency - Inverse Document Frequency) vectoriser, Random forest, ML-KNN(Multi Label K-Nearest Neighbour), Decision tree and Support Vector Machine(SVM). Even yet, there is a paucity of data supporting the effective use of machine learning-based approaches for AMP discovery, and many of these tools are not built to predict putative AMPs' specific functions, such as its antimicrobial activity. As a result, among the seemingly endless array of data mining techniques available for screening peptide sequences for antimicrobial activity, very few are able to perform this work reliably, but with little accuracy and typically no knowledge of potential targets. We hope to provide a user friendly web interface which serves as an AMP(Antimicrobial Peptide) and ABP(Antibacterial Peptide) predicting tool with specific physical, chemical, structural properties as well.

Chapter 1: Introduction

The growing concern over antibiotic resistance highlights the growing interest in antimicrobial peptides, or AMPs. Less than 100 amino acids, or AMPs, are essential components of plant's and animal's host defence mechanisms. [1] Their broad range of activity, rapid death time, and low toxicity make them appealing for use in healthcare research.

Antibiotics are recognised to be successful in combating bacterial infections. Peptide medicines have drawn a lot of attention for their diverse pharmacological uses to treat cancer, autoimmune, hormonal, and metabolic disorders, as well as infectious diseases (viral, fungal, and parasitic). They are also being used to combat multidrug resistance (MDR) and bacterial diseases. Peptide medications are intriguing because they are evolutionary conserved components of innate immunity found in all organisms. [8] While there are many different sources from which these chemicals might be produced, methodical development of therapeutic peptide medicines is important for a number of reasons. [4] It should be known that the threat posed by types of bacteria that are resistant to many drugs and projects, if left unchecked, there might be 10 million deaths per year by 2050. [3] As AMPs are a component of the innate immune system, they are positioned as possible substitutes for antibiotics.

A number of computational tools have been proposed for the identification of AMPs, including CAMP [2], ABP3finder [34], ADAM [2], AntiBP [2], among others. Amino acid composition is a prominent feature for AMPs along with the physical, chemical, structural properties, and many of the tools mentioned above use machine learning techniques including support vector machine (SVM), random forest (RF), and deep learning algorithms involving neural networks. [2]

However, in order to tackle the intricacy involved in feature engineering within machine learning, the study presents a neural network model designed to identify AMPs. By using a multi-label classification with varying labels of anti-gram positive, anti-gram negative and anti-gram variable peptides, outperforms previous models in terms of performance. To turn

amino acids into numerical vectors and extract their semantic information, an embedding layer is used namely TF-IDF (Term Frequency - Inverse Document Frequency).

The project highlights the value of computational tools as an addition to experimental research in drug development, noting that machine learning techniques are useful in a range of biotechnological applications. The study suggests a working pipeline, including data extraction and filtering, classification, regression, implementing 6 different algorithms to rule out the best multi label classifiers, to effectively screen for strong AMPs along with their physico-chemical features and types. This multi-label classification model effectively predicts AMPs from non-AMPs by analysing the user input peptide sequence of any length. Then, it validates from the antibacterial and anti-biofilm peptide dataset curated by us, if the input sequence is an AMP and labels it as anti-gram positive, anti-gram negative or anti-gram variable peptide while demonstrating strong activity against various infections.

Like other peptide medications, the discovery of antimicrobial peptides (AMPs) usually starts with the examination of natural peptides. To find powerful AMPs in larger peptide libraries, other high-throughput experimental techniques—such as the surface-display method—have been utilised, increasing the search space to 0.9 million sequences. [5]

There are computational techniques that have significantly improved AMP sequence optimisation, structural diversity, and therapeutic indexes. These techniques include quantitative structure-activity relationship methods, de novo design, linguistic models, pattern insertion, evolutionary algorithms, and deep generative models with molecular simulations. These approaches are limited in that they can only probe a small area of the large combinatorial molecular space. Additionally, the high production costs linked to larger sequence lengths, proteolytic breakdown, low solubility, and off-target toxicity make it difficult for AMPs generated in silico to advance to clinical trials. [5]

Research Gap

In light of the large combinatorial range of peptide sequences, it is challenging to discover functional peptides in a systematic manner. Throughout history, humans have faced several maladies. There is an upsurge of antibiotics for nearly all bacteria marked for an

immense change and an era of abundant treatment possibilities. However, the advent of multi-drug resistant (MDR) bacteria, which compromised the efficacy of already available antibiotics, made this era short-lived. [6]

Pathogen co-evolution has led to the inevitable emergence of complex infectious illnesses. Antimicrobial resistance is the term used to describe the emergence of infectious illnesses that are resistant to traditional antibiotic treatments due to the selection pressure exerted by antibiotic use. This resistance is present in a broad spectrum of organisms, such as bacteria and it raises the mortality rates worldwide. [7] The World Health Organisation stated in 2024 that the substantial increase in antimicrobial resistance, especially in healthcare settings, has alarmed international health authorities. [8] The use of antibiotics has been identified as the main factor driving the emergence and spread of antimicrobial resistance.

Objectives

- **Data collection** and pre-processing.
- To prepare a small peptide dataset using existing databases with specific information and properties.
- **Enhanced Performance:** To utilise machine learning approach for detection with high accuracy allowing for the intricate correlations between peptide characteristics and antimicrobial action, while also accommodating the intricate nature of biological data. These implementations improve the accuracy and generalisability of the model by limiting overfitting, optimising interpretability, and efficiently employing feature information. This helps identify and create new AMPs with the necessary characteristics.
- **Extending Treatment Options:** By finding new peptides with broad-spectrum activity against a variety of pathogens, including bacteria, viruses, fungi, and parasites, the objective is to diversify the arsenal of antimicrobial medicines that are now accessible.
- **Improving Therapeutic Efficacy:** By using predictive modelling, sequences with strong antimicrobial activity that have the least amount of cytotoxicity and side effects are found, which helps to increase the therapeutic efficacy of AMPs.

- **Optimising Peptide Design:** By comprehending the links between sequence, structure, and function, AMP prediction seeks to optimise the design of therapeutic peptides. This allows for the development of peptides with improved stability, bioavailability, and target specificity.
- The target of this project is to identify and predict a protein sequence given by the user and the algorithm is such written that it returns the Antimicrobial properties, amino acid percentage and also tells about the structural configuration of the protein sequence.
- To emphasise on the necessity of accurate prediction tools and to rule out tools that no longer update their database and give false predictions.

Chapter 2: Literature Review

AMPs as a substitute for traditional antibiotics in the treatment of infectious illnesses

Antimicrobial resistance has emerged as one of the major threats to global public health in recent decades, as the World Health Organisation (WHO) has widely announced in its global report on surveillance. This is primarily due to the widespread use of classical antibiotics in health care systems, animal production, and community settings. Life-threatening consequences may arise from common diseases or mild injuries if this issue is not well managed. Therefore, administering antibiotics as prescribed is insufficient to combat resistant microorganisms. Thus, the development of novel, alternative antibiotics is desperately needed. As part of the innate immune response, the human body creates AMPs, which can be potential candidates for the goal of preventing bacterial infection and/or inhibiting the proliferation of microorganisms due to their numerous advantages. [12]

Combination therapy—which uses conventional antibiotics along with antimicrobial peptides (AMPs)—has become a popular tactic to combat bacterial resistance and improve treatment outcomes. This tactic increases cellular osmolarity imbalance and inhibits repair processes by prolonging the time of bacterial pore opening in addition to defeating resistance. Additionally, AMP-antibiotic combos support a number of other processes, such as the decrease in host cell toxicity and bacterial resistance. [11] Combining AMP with antibiotics creates synergy by focusing on several different, separate bacterial cell processes. Because of this intricacy, the bacteria must experience simultaneous changes in these pathways in order to build resistance, which makes combination therapy an effective tactic against resistance mechanisms. Combinations of AMP and antibiotics are more effective than single medications in preventing the production of biofilms, which goes beyond their antibacterial properties.

Given their effectiveness against organisms that are resistant to many drugs, antimicrobial peptides present a strong argument for addressing the growing problem of antibiotic-resistant diseases. Their principal method of action entails breaking down microbial cell membranes, taking advantage of the unique targets that microbes present. One important way that AMPs suppress or eradicate certain microorganisms is by interfering with cell

membranes. Numerous studies on the combination of AMPs and antibiotics have shown synergistic effects. For example, in both cystic fibrosis patients and in vivo rat models, the colistin sulphate-tobramycin combination shown notable efficiency against *Pseudomonas aeruginosa* biofilms. Another study looked at the combination of human neutrophil peptide (HNP)-1 and rifampicin and isoniazid, two anti-TB medications, to combat *Mycobacterium tuberculosis*. [10] The outcomes showed a significant decrease in the bacterial burden, suggesting that AMPs may improve the efficiency of currently available antibiotics.

Another study concentrating on *Salmonella enterica* serovar Typhimurium showed that ampicillin and cryptdin 2, a Paneth cell antimicrobial peptide, combined to have a potentiated effect on bacterial infection control. [9] When AMPs were used in conjunction with other antimicrobials, the bacterial load was significantly reduced, surpassing that of single-drug therapies.

All things considered, AMP-antibiotic combination therapy is a potential new front in the fight against bacterial infections, providing a variety of strategies to counter resistance and enhance treatment results. The results of the experiments emphasise the possibility of synergistic interactions and the necessity of more research into AMP-antibiotic combinations in various microbiological settings.

Database for AMP extraction and ML models used for prediction

Over a period of time while the AMP research has been conducted, a large number of databases have been developed to predict AMPs. Papers have suggested a DNN model based on the multi-scale convolutional layers to recognise AMPs. The multi-scale convolutional network and the embedding layer are the two primary components of the suggested DNN model. Every amino acid in a peptide sequence is transformed into an embedding vector via the embedding layer. Local features can be captured by the multi-scale convolutional network, and feature selection can be aided by its max pooling layers and convolutional layers with varying filter lengths. This local context-focused methodology has the potential to enhance AMP detection performance.

The multi-scale convolutional network is the most crucial component of our model as, according to the model modification comparisons, the model without it produced inaccurate results. Researchers also used the proposed fusion model and model that was suggested to analyse other datasets, such as the APD3 benchmark dataset, the AIP dataset, and the AMP dataset. The outcomes demonstrate that the fusion model could perform better and that other peptide identification applications could benefit from our suggested model.

With 3146 natural antimicrobial peptides (AMPs) from the six kingdoms (383 bacteriocins/peptide antibiotics from bacteria, 5 from archaea, 8 from protists, 29 from fungi, 250 from plants, and 2463 from animals), 190 predicted, and 314 synthetic AMPs, the Antimicrobial Peptide Database (APD) has 3940 peptides as of January 2024. [21] This database provides more accurate data in comparison to the DBAASP database which has a larger number of peptide sequences but has incorrect data like sequence length and the type of antimicrobial action the peptide performs. [22] The reason for using the APD3 database is well established that it is updated, unique, accurate, however, data was not available to download directly from the database and our dataset is a stand-alone dataset with features extracted manually and data scrapped using Selenium. The addition of non-AMPs has been performed manually from PDB, UniProt and SwissProt.

Action of AMP and ABP on bacterial membrane:

With the objective to increase membrane permeability, inducing cell membrane lysis, and exposing cell content, the cationic (positively charged) AMPs interact with the negatively charged bacterial membrane to execute their antibacterial mechanism. As AMPs reach the cytoplasmic membrane via electrostatic interaction with the microbial membrane, they attach and engage with the plasma membrane's anionic constituents. The capsular polysaccharide and other elements of the bacterial cell wall, such as lipoteichoic acid and peptidoglycan in Gram-positive bacteria and LPS in Gram-negative bacteria, must first be overcome by AMPs. [13,14,15]. To enhance the interaction with the anionic lipid membrane, α -helical AMPs attach themselves to it and change its unstable structure in aqueous solution into an amphiphilic α -helical structure. However, β -sheet peptides have

disulphide bonds and they do not experience a significant conformational change while interacting with the membrane. [16] AMPs are found parallel to one another on the plasma membrane surface at low peptide-lipid ratios. AMPs are vertically oriented and placed into the hydrophobic centre of the membrane as the peptide-lipid ratio rises. Membrane permeability eventually causes internal ions, metabolites, and biosynthesis to ooze out, which results in cell death. [17]

The aggregate model, the barrel-stave model, the toroidal-pore model, and the carpet model are the four main models of membrane-pore creation. When antimicrobial peptides (AMPs) penetrate the phospholipid membrane, their hydrophobic portions combine with the hydrophobic portions of the bilayer located internally, leaving their hydrophilic portions vulnerable. Another way through which AMPs engender bacterial death is by entering the cytoplasm and interacting with components inside the cell. Examples of these interactions include blocking the synthesis of DNA, RNA, and proteins, hindrances to the folding of proteins, lowering the activity of enzymes significantly and hampering the synthesis of cell walls, and encouraging the release of lyases to damage cell structures. [18]

On the other hand, anionic antimicrobial peptides disrupt bacterial cell by membrane dissolution where, certain amino acids like aspartic acid play a crucial role in AMP and target organism membrane interaction. [19]

Hydrophobicity as a parameter for AMP prediction:

Antimicrobial peptides (AMPs), especially those with α -helical structures, rely heavily on hydrophobicity for their mechanism of action. As a component of the innate immune system, AMPs mostly target and damage the membranes of microorganisms. The capacity of AMPs to interact with these membranes is influenced by their hydrophobicity, which in turn impacts both their antimicrobial activity and possible toxicity to host cells. We will examine the complex function of hydrophobicity in AMPs in this in-depth discussion. Hydrophobic fatty acid chains make up a major portion of the lipid bilayer seen in microbial cells. To successfully integrate into this hydrophobic environment, AMPs must have a sufficient degree of hydrophobicity. The integrity of the membrane is compromised by this insertion, which causes cell death and leakage of cellular contents.

AMPs that exhibit more hydrophobicity are more adept at penetrating the hydrophobic core of microbial membranes. For example, up to a specific threshold, a study demonstrated that the peptide V13KL's antimicrobial action is enhanced by increasing its hydrophobicity. [20]

Peptides normally orient themselves once they are close to the membrane in a way that allows their hydrophobic faces to interact with the lipid bilayer and their hydrophilic faces to interact with the aqueous exterior. This orientation is important because it prevents the formation of transmembrane pores or channels, which are known to compromise membrane integrity. [20]

The research also pinpoints a certain hydrophobicity window when antibacterial action is most effective. Within this region, there is minimal hemolytic activity and significant antimicrobial activity displayed by peptides with hydrophobicity.

Over the ideal level of hydrophobicity, additional increases may result in a reduction of antibacterial effectiveness. The reason being, in aquatic conditions, peptide self-association or dimerization increases, resulting in a decrease in the amount of active monomeric peptides that can interact with microbial membranes. There are thirty-eight hydrophobicity scales for peptides (Aboderin, Fauchere, Goldsack, Guy, Jain, Kuhn, KyteDoolittle, Prabhakaran, Rao, Wimley White) that were acquired from diverse sources. The role of Hydrophobic ratios in our project has played a crucial role in streamlining our dataset. We have used the Wimley-White whole residue hydrophobicity scale. Initially, they provide absolute values by taking into account the sidechains and peptide bond contributions. Secondly, the values for the transfer free energy of polypeptides are derived directly from experiments.

Table 1: The following data illustrates the Wimley-White scale used by APD3 database to calculate whole-residue hydrophobicity of the peptide (i.e. the sum of whole-residue free energy of transfer of the peptide from water to POPC (diacylglycerol phospholipid) interface). [21]

Amino Acids	Interface Scale ΔG_{wif} (kcal/mol)
Isoleucine*	-0.31

Leucine*	-0.56
Phenylalanine*	-1.13
Valine*	0.07
Methionine*	-0.23
Proline	0.45
Tryptophan*	-1.85
Histidine	0.17
Threonine	0.14
Glutamic Acid	-0.01
Glutamine	0.58
Cysteine*	-0.24
Tyrosine	-0.94
Alanine*	0.17
Serine	0.13
Asparagine	0.42
Aspartic Acid	-0.07
Glycine	0.01
Arginine	0.81
Lysine	0.99

We have utilised the above mentioned Wimley White scale. However, our model calculates hydrophobic ratio on the basis of frequency of occurrence of the hydrophobic peptides starred above. Alanine (Ala, A), valine (Val, V), leucine (Leu, L), isoleucine (Ile, I), methionine (Met, M), phenylalanine (Phe, F), tryptophan (Trp, W) and cysteine (Cys, C) are among the hydrophobic amino acids. The hydrophobic core of proteins, which is separated from the polar solvent, is usually formed by these residues. These hydrophobic cores are due to Van Der Waals interactions that are crucial for stabilising the structure and are facilitated by closely packed side chains.

Net Charge use for prediction

Antimicrobial peptides have a positive net charge (of at least +2) at pH 6 to 7, which provides binding specificity to the negatively charged bacterial membranes through electrostatic interactions. Certain side chains of amino acids can impart electric charge to the proteins under specific pH values. The sum of the charges of each amino acid is called net charge.

Overview of ML Approach for Prediction

Classification in Binary aims to forecast just one functional class while making use of binary classifiers to streamline the correspondence between labels and features. Multiple Label Labelling aims to predict several different functional classes at once by using two primary techniques to employ multi-label classifiers: binary relevance and algorithm adaptation. [25]

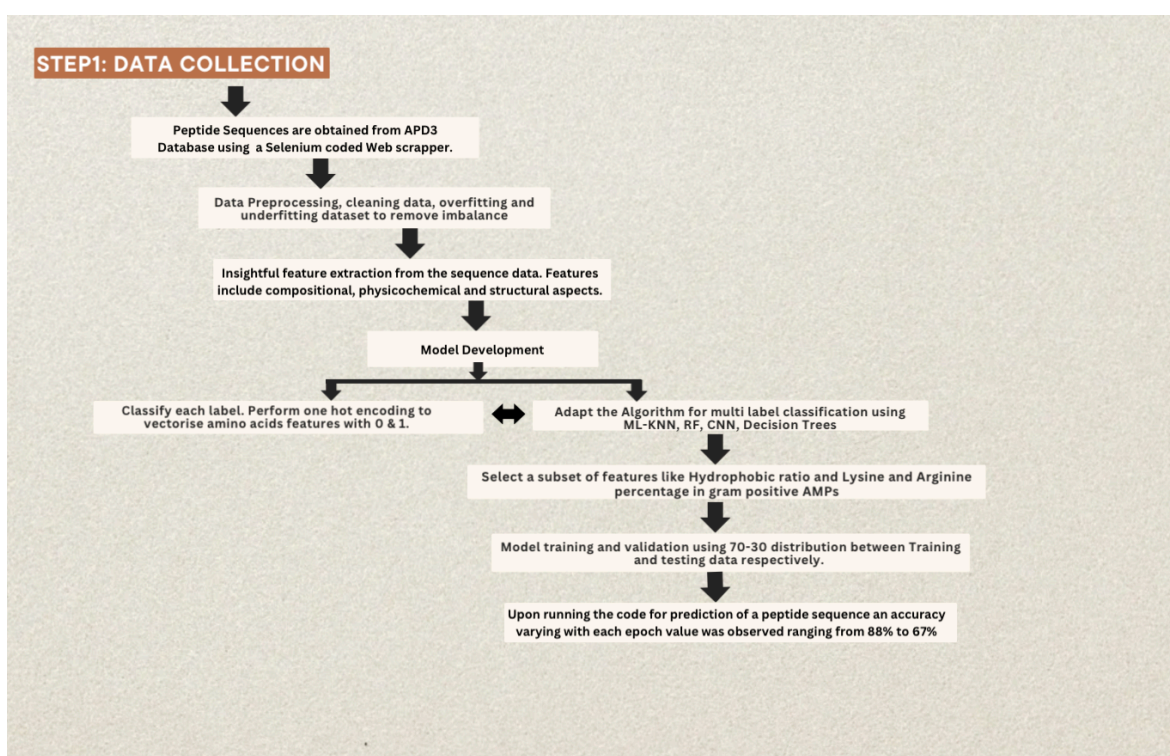


Figure 1: The methodical methodology used in the study to forecast AMP functional classes is visually represented by this flowchart, which highlights the phases from acquiring data to model rebuilding and performance improvement.

Classification of AMPs with reference to APD3 (Antimicrobial Peptide Database 3)

The ADP3 database's [21] statistics show that there are eighteen categories into which AMP activity can be classified. Antibacterial, antiviral, antifungal, antiparasitic, antibiofilm, antidiabetic, antitoxin, antiHIV, and anti-cancer peptides are some ways to summarise these groups. Most AMPs are derived from animals followed by bacteria and plants. Since, drug resistance is developed by the bacterial infection negligent treatment it becomes crucial to identify, predict and synthesize Antibacterial peptides. Moreover, the emphasis should be more on ESKAPE pathogens (Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, and Enterobacter) species are among the bacteria that make up the ESKAPE pathogens category. These bacteria can be either Gram-positive or Gram-negative, in addition to the patient's disregard for the antibiotic regimen, these pathogens/superbugs have the ability to mutate quickly, which aids in the development of resistance against antimicrobials. [25] However, multicellular organisms naturally manufacture AMPs, which serve as their first line of defence against dangerous bacteria during infections. AMPs are cationic in nature, amphiphilic in nature, and generally tiny (10–50 amino acids). Humans are naturally immune to microbial infections, including those caused by lysozyme, which is released by the nasal mucosa and functions as a bacteriolytic. Widely recognised for their distinctive size, polypeptides can exist in primary, secondary, tertiary, or quaternary conformations, adding to their flexibility and amphiphilic nature. Additionally, their surface charge is complementary to the surface charge of bacteria on their cell membranes giving them an upper hand against the microbes. [26,27]

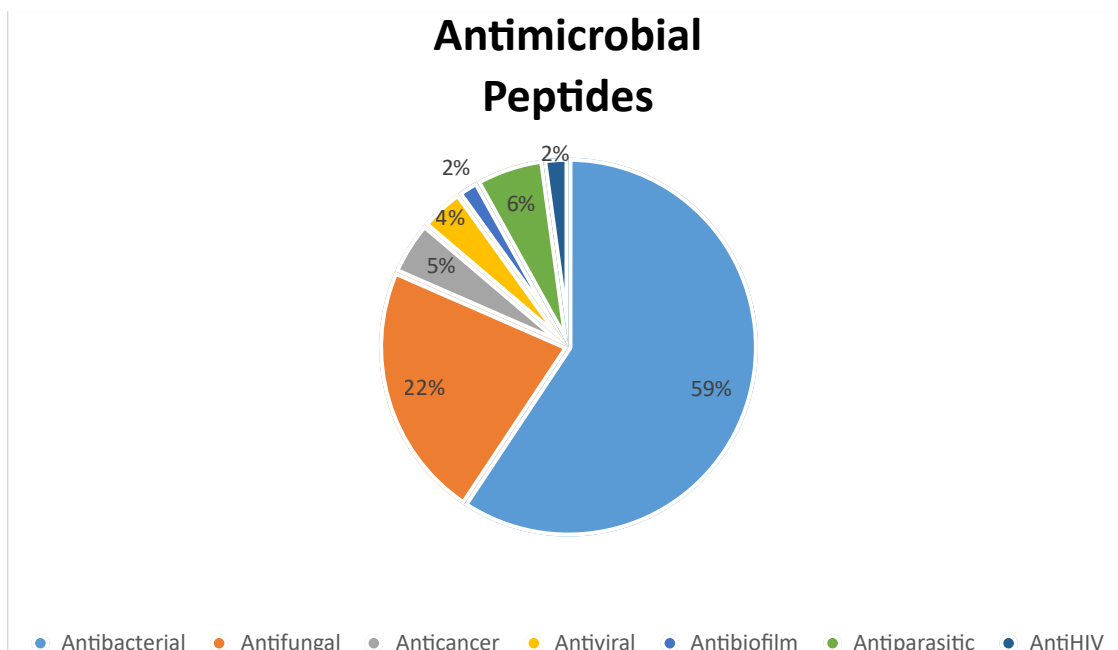


Figure 2: AMP sequences availability in APD3 updated database. Most of these peptides shared properties for a larger category on which we have trained our model. Out of 1400 Antifungal AMPs 1200 are active against both gram-positive, gram-negative and gram variable bacteria. 244 out of 294 anticancer AMPs are active against both gram-positive and gram-negative. 162/244 antiviral and 104/116 antibiofilm AMPs show antimicrobial activity against gram bacteria. [21]

Chapter 3: Materials and Methods

We utilised precompiled anti-microbial peptide databases and to test the results of this tool we decided to run it multiple times on the following parameters:

1. **Host System:** Apple MacBook Pro
2. **Processor:** Apple M1
3. **RAM:** 8 GB - 128-bit LPDDR4X SDRAM
4. **SSD:** 512 GB
5. **Browser:** Safari
6. **Cloud Platform:** Google Colab (Free Tier)
7. **Dataset:** Antimicrobial Peptide Database (University of Nebraska) [6]

A. Curating the right dataset

When it comes to machine learning models' ability to predict Antimicrobial Peptide (AMP) sequences, the calibre of the training set is a crucial factor. Since this training set is the main source of data used in the model's learning process, its content and quality are very important. AMP databases are usually the source of the AMP sequences that are used for training. The methods used for data gathering, curation, and maintenance differ greatly among these databases. Large Antimicrobial Peptide and Protein Database (LAMP2), Antimicrobial Peptide Database (APD), Collection of Antimicrobial Peptides (CAMP), dbAMP, Database of Antimicrobial Activity and Structure of Peptides (DBAASP), and dbAMP are important AMP databases. [22]

Distribution of Peptide Lengths: Databases differ in how they distribute peptide lengths. For instance, 90% of the entries in APD, which is primarily composed of natural AMPs, are less than 60 amino acids, whereas 1,990 sequences in CAMP are longer than 150 residues.

Composition of Training Sets: APD and CAMP have a substantial overlap, as a result of CAMP incorporating several natural AMPs from APD. A hybrid technique has been used in recent studies to improve the diversity and quantity of the training set by combining

sequences from other databases. Prediction results are significantly influenced by the amount of the positive dataset.

Summary Statistics of Numerical Features:			
	hydrophobic	netcharge	length
count	3188.000000	3188.000000	3188.000000
mean	46.881744	4.385508	25.704831
std	11.499477	2.885277	10.551965
min	20.000000	-7.000000	8.000000
25%	38.000000	2.000000	17.000000
50%	46.000000	4.000000	24.000000
75%	55.000000	6.000000	34.000000
max	93.000000	22.000000	50.000000

Figure 3: Summary of statistics of features of sequences present in the dataset.

B. Testing Dataset and the complexities

Peptides with experimentally confirmed lack of antibacterial activity should make up an optimal negative dataset. Unfortunately, there is a dearth of verified non-AMP sequences in public databases as a result of the infrequent publication of these non-AMP sequences. The quality of training sets would be greatly improved by establishing a special database for non-AMP sequences while motivating scholars to contribute to it.

In response to the lack of verified non-AMP sequences, bioinformaticians have created negative datasets using over 200 million sequences from protein sequence databases like UniProt, PDB and SwissProt. Since these sequences are not classified as biohazardous, secretory, or penetrating through bacterial membranes, they offer an almost limitless reservoir for negative datasets.

Nevertheless, there could be problems with this strategy, like sequences with unidentified antibacterial qualities being accidentally included. To detect and reduce this kind of contamination, training models with various sets of positive and negative datasets can be helpful. Therefore, for this project I curated most of the non-AMP data from protein databases by applying the same condition and features for sequences across all organisms. It is generally recommended to use a balanced training set that has an equal representation

of AMP and non-AMP sequences. Sequence identity, length, activity, and other factors are taken into consideration while selecting AMP sequences from one or more AMP databases.

We utilised Machine Learning/Deep Learning approaches and principles for accurate prediction of anti-microbial property of small peptides:

- **Sequence Understanding:** Antimicrobial peptide (AMP) prediction uses machine learning Long Short-Term Memory (LSTM) models, Random Forrester Classifiers, ML-KNN, Decision trees because of their superior ability to comprehend and interpret sequential input is crucial to enable auto feature extraction as the model self teaches the features of the peptides.
- RF is a member of the ensemble techniques family, which uses decision trees as foundation classifiers. In a recent comparison, RF and deep learning techniques demonstrated similar modelling performance for AMP datasets. [28,29]
- Because of their interpretability, flexibility with regard to data assumptions, tolerance to outliers, and ability to handle complex and non-linear interactions, decision trees are an extremely useful tool for predicting antimicrobial peptides. Their special suitability for AMP research stems from their capacity to recognise important aspects and their incorporation into ensemble approaches. [30]
- Through the application of kernel functions (such as radial basis function and polynomial), which translate the input features into higher-dimensional spaces where data is separable, SVMs are able to handle complicated, non-linear interactions in biological samples. In AMP prediction, where peptide sequences and their antimicrobial activities might have complex and non-linear relationships, this is very helpful.
- Moreover, SVMs have a lower tendency to over fit, particularly on short datasets like those used in AMP investigations. To improve the model's generalizability to new data, their regularisation parameters manage the trade-off between obtaining a low training error and a low testing error. This is essential to guarantee the prediction model's continued accuracy and robustness when used with new peptide sequences. [31]

- Because of how well it handles multi-label classification problems, Multi-Label k-Nearest Neighbour (ML-KNN) is an important technique for the estimation of antimicrobial peptides (AMPs). To predict multiple functional labels for each peptide—a crucial feature for AMPs, which frequently display multiple activities like antibacterial, antifungal, antiviral, and anticancer properties—ML-KNN is designed to differ from conventional approach to classification that assign a single label to each instance.
- In order to capture the intricate linkages and interactions between various peptide properties, our method makes use of the local information supplied by the nearest neighbours. With ML-KNN, many capabilities may be reliably predicted by taking into account the label distribution across the neighbours, giving each AMP a detailed profile.
- Word Embedding: The numpy and Tensorflow vectoriser TF-IDF (Term Frequency Information Document Frequency of records). This contributes to the better understanding of the peptide sequence by the model by helping to capture semantic information and links between various amino acids.
- Feature Learning: From sequential data, they examine similarity metrics and label distribution among closest neighbours, ML-KNN (Multi-Label k-Nearest Neighbour) infers implicitly the significance of features. Using bagging to choose features, Random Forest determines the value of each feature by evaluating how it reduces impurities across decision trees. Information gain or the Gini index are used by decision trees for hierarchical feature selection. Less significant characteristics may be pruned to increase relevance. Vital support vectors for categorisation are highlighted by SVM (Support Vector Machines), which convert characteristics into higher-dimensional spaces using kernel approaches. [32]

Effective feature learning improves the predictive accuracy of models in AMP prediction and other bioinformatics tasks. ML-KNN's emphasis on analogy and label distribution, Random Forest's ensemble approach, Decision Trees' hierarchical selection, and SVM's emphasis on support vectors all play a part in this. In addition to preventing overfitting and enhancing model interpretability, these techniques take into account the richness and

diversity of biological data and offer insights into the connections between peptide characteristics and antimicrobial action.

Properties and Implementation of curated dataset

Initially, we collected data by scraping many public online sources in order to generate a comprehensive and diverse dataset for our project. After acquiring the data, we employed techniques to parse and pre-process the protein sequences in order to prepare them for feature extraction. Key properties, such as the hydrophobicity ratio, were calculated with established methods, providing useful details regarding the biological properties of the peptides.

Bacteria type	Average hydrophobic ratio	Range	Net Charge at pH 6.5	Range	Average length of peptide
Gram positive	46.5%	20%-84%	+3.5	-9 to12	30.7
Gram negative	39.5%	20%-87%	+4.7	-6to40	36.5
Gram variable/ independent	45.7%	20%-93%	+4.9	-12 to30	30.0

Table 2: The following data was calculated after tabulation in excel. The purpose was to gain some insights about these properties and use them for accurate predictions. However, the data was massive and inconclusive, it helped us in hyper tuning the right parameters, analysing an in between range to exhibit these properties for laying down the right conditions. The data led to the execution of setting the hydrophobic values at a minimum 20% for a peptide input of any length betwixt 8 and 50.

Calculation of Hydrophobic residues

A python code to provide in-built functionality of calculating the features including but not limited to hydrophobic ratio of the input peptide sequence was written. A measure of the

hydrophobicity of a peptide sequence explains about the attacking mechanism of AMPs. The tendency of non-polar molecules to form aggregates in order to reduce contact surface with polar molecules like water. The hydrophobicity of a peptide sequence can be calculated using different scientific indexes expanding and using different properties to determine hydrophobicity. Furthermore, the demand for more details implies a certain level of proficiency from the user, implying a thorough comprehension of peptide properties and categorization labels.

✓ Calculating Hydrophobic Ratio

```
[ ] def calculate_hydrophobic_ratio(peptide_sequence):  
    # List of hydrophobic residues based on general classification  
    hydrophobic_residues = {'A', 'C', 'F', 'I', 'L', 'M', 'V', 'W'}  
  
    # Initialize counters for total and hydrophobic residues  
    total_residues = len(peptide_sequence)  
    hydrophobic_count = 0  
  
    # Iterate over each residue in the peptide sequence  
    for residue in peptide_sequence:  
        # Convert the residue to uppercase to ensure case-insensitive matching  
        residue = residue.upper()  
  
        # Check if the residue is in the list of hydrophobic residues  
        if residue in hydrophobic_residues:  
            hydrophobic_count += 1  
  
    # Calculate the hydrophobic ratio percentage  
    hydrophobic_ratio_percentage = (hydrophobic_count / total_residues) * 100  
  
    return hydrophobic_ratio_percentage
```

Figure 4: The above code creates a Python function called calculate hydrophobic ratio. The peptide sequence is represented as a string by the function, which accepts a single parameter called peptide sequence. The function defines a collection called hydrophobic residues, which is made up of amino acid residues that have been categorised as hydrophobic using a general categorisation system. These amino acids, as we have stated in our code are, 'A', 'C', 'F', 'I', 'L', 'M', 'V' and 'W'. Two counters are also initialised: hydrophobic count, which keeps track of the number of hydrophobic residues encountered, and total residues, which stores the total amount of residues in the peptide sequence. Every residue in the peptide sequence is iterated over by the function. First of all, it makes all

residues uppercase to provide case-insensitive matching. Next, it determines whether the residue falls within the hydrophobic residues category. The hydrophobic count is increased if the residue has a hydrophobic character. Following the processing of each residue, the function multiplies the result by 100 and divides the hydrophobic count by the total number of residues to determine the hydrophobic ratio percentage. Use the round function to round the result to the closest integer.

Scrapping for data extraction from APD3 database

```
scraper.py > ...
1  from selenium import webdriver
2  from selenium.webdriver.common.by import By
3  import csv
4  from csv import writer
5
6  driver = webdriver.Chrome()
7
8  driver.get("https://aps.unmc.edu/home")
9  driver.implicitly_wait(2)
10 driver.find_element(By.CLASS_NAME, "link_button").click()
11
12 driver.implicitly_wait(2)
13 peptide_ID_elements = driver.find_elements(By.CLASS_NAME, "link_button")
14
15 peptides = []
16 for p in range(len(peptide_ID_elements)):
17     # for p in range(0,500):
18     driver.implicitly_wait(2)
19     peptide_ID_ele = driver.find_elements(By.CLASS_NAME, "link_button")
20     current_peptide_ID = peptide_ID_ele[p].text
21     peptide_ID_ele[p].click()
22
23     # Name/Class
24     if (driver.find_element(By.XPATH, "/html/body/div[2]/div[2]/table/tbody/tr[2]/td[2]")):
25         peptide_name_class = driver.find_element(By.XPATH, "/html/body/div[2]/div[2]/table/tbody/tr[2]/td[2]").text
26     else:
27         peptide_name_class = " "
28
29     # source
30     if (driver.find_element(By.XPATH, "/html/body/div[2]/div[2]/table/tbody/tr[3]/td[2]")):
31         peptide_source = driver.find_element(By.XPATH, "/html/body/div[2]/div[2]/table/tbody/tr[3]/td[2]").text
32     else:
33         peptide_source = " "
34
35     # additional_info
36     if (driver.find_element(By.XPATH, "/html/body/div[2]/div[2]/table/tbody/tr[14]/td[2]")):
37         peptide_additional_info = driver.find_element(By.XPATH, "/html/body/div[2]/div[2]/table/tbody/tr[14]/td[2]").text
38         activities = peptide_additional_info.find_elements(By.TAG_NAME, "p")
39         for activity in activities:
40             if activity.text != "" and activity.text.split()[0] == "Activity:":
41                 peptide_additional_info_activity = activity.text.split(" ", 1)[1]
42                 break
43         else:
44             peptide_additional_info_activity = " "
45
46     # activity
47     if (driver.find_element(By.XPATH, "/html/body/div[2]/div[2]/table/tbody/tr[12]/td[2]")):
48         peptide_activity = driver.find_element(By.XPATH, "/html/body/div[2]/div[2]/table/tbody/tr[12]/td[2]").text
```

Figure 5: The script streamlines the procedure of gathering data on antimicrobial peptides from websites and converting it into a CSV file. It opens a web driver, sets up the destination page, and loops through the peptide entries. It pulls specified information from each entry, manages any exceptions, and appends the data to a CSV file. By repeating this method for a predetermined range of entries, a sizable dataset can be assembled for additional study or analysis.

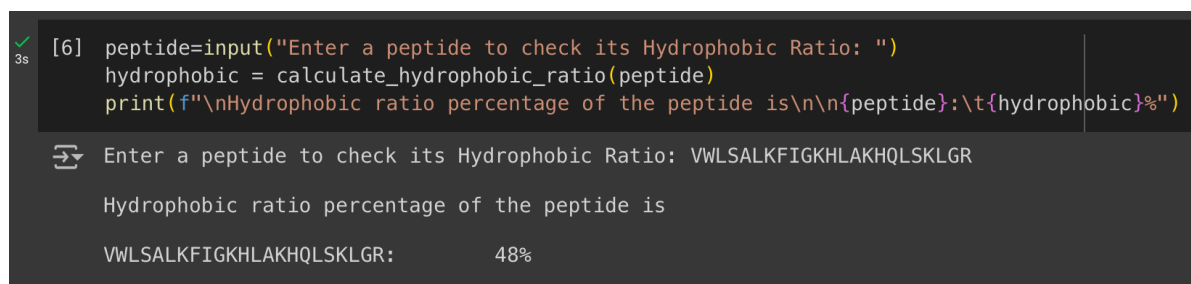
Methodologies and results followed by future prospect of this project which may be utilised for implementation in a week's time to design a user interface for prediction where, user input is taken and our machine learning algorithm running in the backend predicts the most accurate answer.

Chapter 4: Results

We prepared the dataset as well as the tools for prediction of hydrophobicity and net charge following which adequate results were seen.

A. Hydrophobicity Calculations

When choosing the ideal operating parameters for a hydrophilic-ionic catalyst (HIC) process, one crucial aspect to consider is the hydrophobicity of the biomolecules. Though some of them are also commonly found on the surface of proteins, the native structure of proteins generally contains the highest concentration of hydrophobic amino acids in their internal core.



```
[6] peptide=input("Enter a peptide to check its Hydrophobic Ratio: ")
hydrophobic = calculate_hydrophobic_ratio(peptide)
print(f"\nHydrophobic ratio percentage of the peptide is\n\n{peptide}:\t{hydrophobic}%")

Enter a peptide to check its Hydrophobic Ratio: VWLSALKFIGKHLAKHQLSKLGR

Hydrophobic ratio percentage of the peptide is

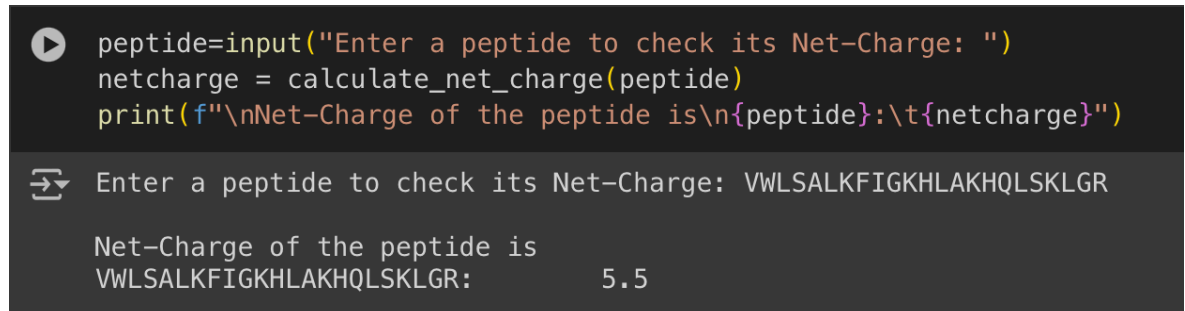
VWLSALKFIGKHLAKHQLSKLGR:          48%
```

Figure 6: Using established a pre-determined set of amino acids, our script efficiently determined the hydrophobic ratio for every protein sequence under investigation. These scales assign a hydrophobicity rating to each amino acid residue based on whether the residue is more likely to be exposed to the surrounding solvent or buried in the protein core.

B. Calculation of Net Charge

A Python tool was developed to calculate the net charge of the user input peptide sequence. It is crucial to understand that the behaviour and functionality of peptides, particularly in relation to their interactions with other molecules and overall structural stability plays a pivotal role in membrane interaction between AMPs and microbes. The script calculates the net charge for each peptide sequence by adding together all of the charges of the amino acid residues at a specific pH. The pH was experimentally and after consulting other net charge scales and estimation tools, the pH It had to be iteratively derived because of the

limitation that net charge values only near this pH were matching those of the dataset. Hence, it was experimentally achieved to avoid inconsistency.



```
▶ peptide=input("Enter a peptide to check its Net-Charge: ")
netcharge = calculate_net_charge(peptide)
print(f"\nNet-Charge of the peptide is\n{peptide}:\t{netcharge}")

↵ Enter a peptide to check its Net-Charge: VWLSALKFIGKHLAKHQLSKLGR

Net-Charge of the peptide is
VWLSALKFIGKHLAKHQLSKLGR:      5.5
```

Figure 7: We compared the outcomes generated by this script with those from tests and widely used computational tools in order to verify the accuracy of our net charge computations. The high degree of agreement between the two comparisons suggests that our method produces accurate and dependable net charge estimates for protein sequences.

C. Machine Learning Approaches and Predictions

We employed the algorithm to determine the net charge and hydrophobic ratio of a given peptide, which we could approximate rather well, as features for our machine learning models. We implemented Random Forest, Decision Tree, Support Vector Machine (SVM), and Multi-Label K-Nearest Neighbours (KNN). The results clarified the aspects we executed and pointed out areas that could be enhanced in future studies. In this section, we discuss and compare the results obtained.

It was necessary to develop a fundamental model that could assess whether a certain peptide sequence contains anti-microbial capabilities before exploring the prediction of sub-classes inside a predetermined anti-microbial peptide sequence. In order to do this, we classified peptides as either non-anti-microbial or anti-microbial using a Logistic Regression model. The Logistic Regression model performed admirably, producing results that were acceptable for our categorization assignment (Fig. 8).

Accuracy: 97.07%					
	precision	recall	f1-score	support	
0	0.96	0.97	0.96	409	
1	0.98	0.97	0.98	648	
accuracy			0.97	1057	
macro avg	0.97	0.97	0.97	1057	
weighted avg	0.97	0.97	0.97	1057	

Figure 8: Classification report of logistic regression model. The categorisation report offers comprehensive insights into the accuracy and performance indicators of the model. For this particular issue statement, our Logistic Regression model remarkably obtained an accuracy of 97.07%.

This higher degree of precision in the above results highlights a good Logistic Regression method which performs to determine, whether the input peptide sequences are anti-microbial or not. The data such obtained is comprehensive as the non-AMP distinguishing feature provides insights not only it's hydrophobic content but also its therapeutic properties.

Further we checked our model for different peptides for their anti-microbial properties. Figure 9 shows a clear result of accurate predictions made by our model for randomly selected peptides. It is clearly proved from results in the form of figures (9-14) that the accuracy of our model stands true to its predictions. Each result is described in detail for their anti-microbial property and hence proving our tool as a prediction model.

```

Predictions for sequence MDYGVKASHFTPVGNNALLYTLKQGLGEKWNPELRQAWVDTFRVVATVMKAHSFSH:
Non Anti-Microbial Peptide
-----

Predictions for sequence KGRCFGPSICCGDELGCFVGTAEALLCREENYLPSPCQSGQKPCGSGGRCAAAGICCSF
Non Anti-Microbial Peptide
-----

Predictions for sequence GSKKPVPPIIYCNRRRTGKCQRM:
Anti-Microbial Peptide

Hydrophobic Ratio: 29%
Net-Charge: 6.0

KNN predicted: Anti-Gram+ Anti-Gram-
DT predicted: Anti-Gram+ Anti-Gram-
RF predicted: Anti-Gram+ Anti-Gram-
Voting predicted: Anti-Gram+ Anti-Gram-
-----

Predictions for sequence GKWMSLLKHILK:
Anti-Microbial Peptide

Hydrophobic Ratio: 50%
Net-Charge: 3.2

KNN predicted: Anti-Gram-
DT predicted: Anti-Gram+
RF predicted: Anti-Gram+
Voting predicted: Anti-Gram+

```

Figure 9: The first 2 sequences are non AMPs. 1) First sequence was obtained from NCBI and is a globin domain-containing protein of the *Pseudoalteromonas* species. [35] 2) The second sequence is oxytocin-neurophysin 1 precursor another peptide predicted as non AMP and acts as a muscle contractor during parturition. The algorithm further predicts only those sequences as antibacterial those who have been predicted as AMPs. 3) The sequence predicted as both AGP and AGN is of a chain A protein, Thanatin which is a pathogen-inducible single-disulfide-bond-containing β -hairpin AMP which was first isolated from the insect *Podisus maculiventris*. [42] [41] 4) Halcitine, an anti-gram negative peptide correctly predicted by KNN. [36]

```
Predictions for sequence GILDAIKAIKAAG:
Anti-Microbial Peptide

Hydrophobic Ratio: 64%
Net-Charge: 1.0

KNN predicted: Anti-Gram+
DT predicted: Anti-Gram+
RF predicted: Anti-Gram+
Voting predicted: Anti-Gram+

-----

Predictions for sequence ITSISLCTPGCKTGALMGCNMKTATCHCSIHVSK:
Anti-Microbial Peptide

Hydrophobic Ratio: 44%
Net-Charge: 3.4

KNN predicted: Anti-Gram-
DT predicted: Anti-Gram+
RF predicted: Anti-Gram+
Voting predicted: Anti-Gram+

-----
```

Figure 10: 1st sequence is of Hylaseptin an alanine rich naturally occurring antigram positive peptide. 2nd sequence is the strains of *Lactococcus lactis* sub specie *lactis* (*Lactis lactis*) generate nisin, a member of class I bacteriocins known as lantibiotics and as predicted is active against both gram -ve and gram +ve bacteria. [45]

```
print_label_function('GGAGHVPEYFVGIGTPISFYG')

Hydrophobic Ratio: 33%
Net-Charge: -0.8

KNN predicted: Anti-Gram-
```

Figure 11: Glycine rich antigram negative peptide Microcin, a class 1 bacteriocin.

```

Predictions for sequence MKILYLLFAFLFLAFLSEPGNAYKQCHKKGGHCFPKEKICLPPSSDFGKMDCRWRWKCKCKGSGK:
Anti-Microbial Peptide

Hydrophobic Ratio: 43%
Net-Charge: 9.4

KNN predicted: Anti-Gram+ Anti-Gram-
DT predicted: Anti-Gram+ Anti-Gram-
RF predicted: Anti-Gram+ Anti-Gram-
Voting predicted: Anti-Gram+ Anti-Gram-

```

Figure 12: Crotamine, a gram variable peptide was picked up from UniProt, interacts with heparan sulphate proteoglycans and accumulates in lysosomal vesicles in order to enter the cell by clathrin-mediated endocytosis. The contents of these vesicles, when they burst.

```

Predictions for sequence QCIGNGGRCNENVGPPYCCSGFCLRQPGQGYGYCKNR:
Anti-Microbial Peptide

Hydrophobic Ratio: 27%
Net-Charge: 2.9

KNN predicted: Anti-Gram-
DT predicted: Anti-Gram+
RF predicted: Anti-Gram+
Voting predicted: Anti-Gram+

```

Figure 13: Is a plant derived AMP named Mirabilis jalapa AMP 1. It is non-toxic to grown human cells and Gram-negative bacteria, however it has antifungal activity and action against two tested Gram-positive bacteria. [40]

```

VFIDILDKVENAHNAQVIGIGFAKPFKELINPK: Anti-Microbial Peptide
ILPEHYPIVGTCLLQAIREVLAGAETATDEVIAAWGEAYQQLADILIGAHEVYENIAAAPGGWRGGRMFVKAKTPESDEITSFYLEPLDGQPVIHAKAGQYIG: Non Anti-Microbial Peptide
MNNKQKALKVSTVAVLKSNGADLTQYFYNRMFNHNPENKLLT: Non Anti-Microbial Peptide
MKKINGWIVVALLAVTTVGAATAIQTNNVADSPGQFQVAQKQMY: Non Anti-Microbial Peptide
INLKALAAKAKIL: Anti-Microbial Peptide
GSKKPVPYIYCNRRGTGKQRM: Anti-Microbial Peptide
MAFLKKSFLVFLFGIVLSVCEEEKREGEKEEKREKEEENEDGNEEHKEKRFGLGAILKIGHALAKTVLPMVTNAFKPKQ: Anti-Microbial Peptide
MRISEARKLKEGDIVITPHGPLETVCISSEFNSPLGRNTIVYVKGKTDNGGLMKFHKELKLEGSHENR: Non Anti-Microbial Peptide
MIRCLKVLAIIFSICALFQIHCSLHPENAPLVRPKRMTPFWRGVSRRVPGAPCRDNSECFGTGVCRNKQCSLRILQE: Anti-Microbial Peptide
MILTAVLGGARRPETRARRVAETVRLAALNKRANVDRE: Non Anti-Microbial Peptide

```

Figure 14: The above prediction results were made using our AMP and ABP predicting machine learning algorithm. The peptides were obtained from NCBI, UniProt, APD3

database manually by applying necessary filters like non AMPs, peptides active against gram negative/ positive bacteria, etc.

We combined four different methods with one hybrid strategy to get an accurate prediction model. Alongside the independent techniques, additionally, we employed a hybrid approach that incorporated an ensemble voting system. A variety of performance standards were used to evaluate the efficacy of various methods.

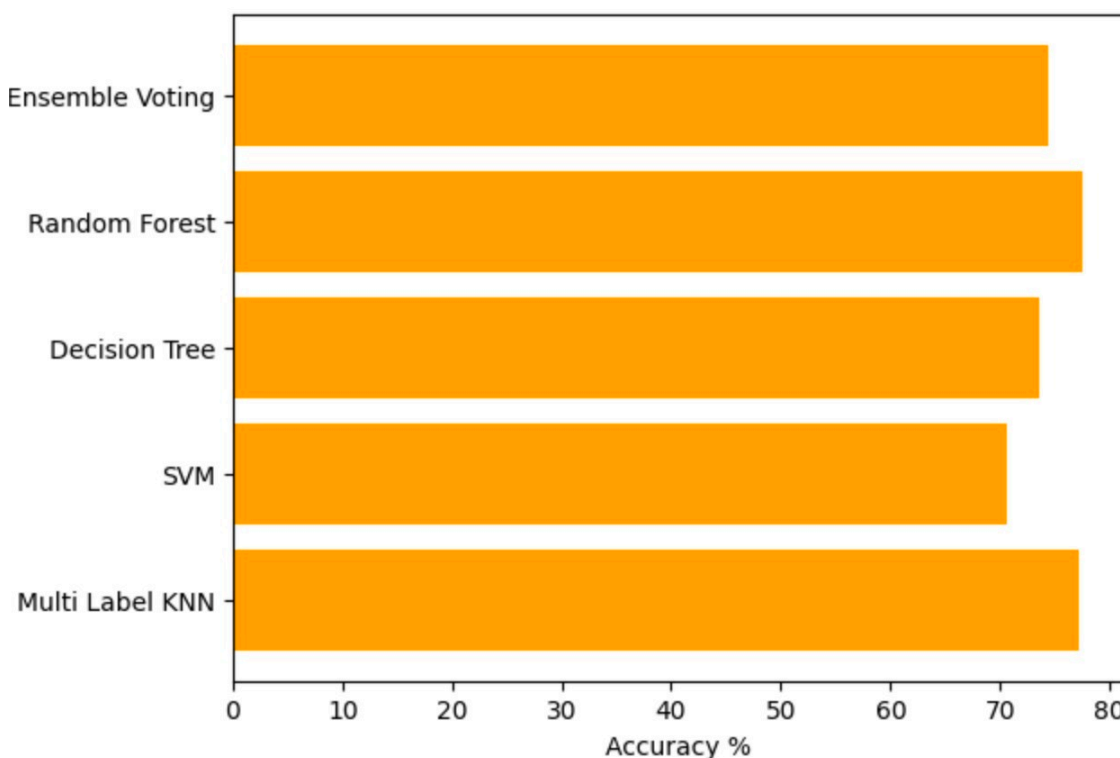


Figure 15: Accuracy Comparison shows that Multi-Label KNN and Random Forest fared better in terms of accuracy than the other models. However, the Support Vector Machine's accuracy was a little bit worse. Even while Multi-Label KNN and Random Forest did rather well, the absolute accuracy percentages were still below the 90% threshold, which is typically regarded as good performance in predictive modelling.

To enhance the useful prediction abilities of these models, we developed an ensemble voting classifier. This technique combines the predictions from all the different models and uses the classifiers' majority vote to decide the final labels. The ensemble voting classifier aims to improve the overall prediction accuracy and robustness of the system by using the strengths of each model.

It was demonstrated that the ensemble technique has the ability to reduce the drawbacks of individual classifiers and generate a prediction result that is more reliable. The ensemble voting classifier reduces the likelihood of errors associated with any one model by aggregating the predictions of several models, improving the prediction model's overall performance and reliability. Because it balances the many biases and fluctuations present in each model, this hybrid approach is crucial in scenarios where reliability and high accuracy are needed.

```
Input the sequence to be checked:> LFKLLGKIIHHVGNFVHGFSHVF
Hydrophobic Ratio: 52%
Net-Charge: 3.0

KNN predicted: Anti-Gram-
SVM predicted: Anti-Gram+
DT predicted: Anti-Gram+
RF predicted: Anti-Gram+
Voting predicted: Anti-Gram+
```

Figure 16: Accurate prediction using voting system. This shows the results that we obtained when we tried to predict the features of a peptide sequence.

After employing all the above stated algorithms, we can see how three out of four independent algorithms predicted that the peptide sequence *LFKLLGKIIHHVGNFVHGFSHVF* is supposed to be Anti-Gram+ only. However, we can also observe that KNN model predicted it to be an Anti-Gram- only. KNN predicts incorrectly here, but to a layman, how will the decision be made of trusting the right output. That's what is solved by our ensemble voting mechanism. The voting classifier determines which labels are said to be true by a majority of the given classifiers, and consequently results in a better probability of predicting the right labels.

```

30/30 [=====]
Test Accuracy: 0.8840125203132629

Test Accuracy: 0.8578892350196838

```

Figure 17: Remarkable performance of Machine Learning algorithms that were tested using other techniques. A range of accuracies were seen, as shown in Figure 10, with a couple of screenshots, with the highest accuracy being achieved at over 88%. A range of accuracies were seen, as shown with a couple of screenshots, with the highest accuracy being achieved at over 88.

Predictions		Actual Label
[1 1]	✓	[1 1]
[1 1]	✓	[1 1]
[1 1]	✓	[1 1]
[1 1]	×	[1 0]
[1 0]	✓	[1 0]
[1 1]	✓	[1 1]
[1 1]	✓	[1 1]
[1 0]	×	[1 1]
[1 1]	✓	[1 1]
[1 1]	×	[0 1]
[0 1]	×	[1 1]
[1 1]	✓	[1 1]
[1 1]	✓	[1 1]
[1 0]	✓	[1 0]
[1 1]	✓	[1 1]
[1 1]	✓	[1 1]
[1 1]	✓	[1 1]
[1 1]	✓	[1 1]
[1 1]	✓	[1 1]
[1 1]	×	[0 1]
[1 1]	✓	[1 1]
[1 1]	✓	[1 1]
[1 0]	✓	[1 0]
[1 1]	✓	[1 1]

Figure 18: Some prediction snippets made by our model. The 1,1 label stands for AMP active against both gram-positive and gram negative bacteria. 0,1 the peptide is functional against gram negative only and 1,0 hinders the gram positive bacterial membrane.

Chapter 5: Discussions

Positively charged, short-chain molecules called antimicrobial peptides work against a variety of pathogens by engaging with the intended cell components through several ways. In contrast to traditional therapies, bacterial resistance to AMPs is more complex to evolve because of their many modes of action on the membrane. AMPs are therefore a desirable substitute for fighting bacteria that are resistant. However, there are certain drawbacks to using AMPs produced from natural sources, such as limited stability, high toxicity, and salt tolerance, which restrict their therapeutic uses.

The impact of the peptides' physico-chemical characteristics on the stability and activity of AMPs is better understood thanks to computational research on AMPs. The aforementioned challenges can now be overcome and peptides with broad-spectrum activity and good stability can be designed with the aid of computational techniques in the research of AMPs.

In this work, a machine learning-based algorithm for predicting the peptides active against Gram-positive, Gram-negative, Gram-variable bacteria and AMP/non-AMPs were independently established for the first time. The cell-surface structures of Gram-positive and Gram-negative bacteria are known to differ from one another which in turn affects the way peptide interact with them.

While going through databases for AMP sequences, I often came across error in prediction and these peptides were pre-trained and their domains were pre-specified. For instance, the DBAASP database does not offer the filtering of data precisely and stores incorrect amino acid length was not updated. The APD3 database has a property calculating tool which calculates the hydrophobicity, net charge but does not predict any peptide outside its database. However, my multi label classification approach promises predictions with accurate data on hydrophobicity and net charge as well. The various machine learning models deployed extract features from their understanding of dataset trained and performs multi label classification with every 1 in 4 random predictions being correct. Compiling the dataset was the most challenging work as it requires balance. The number of peptide

sequences for both AGN and AGP peptides is less in comparison to gram variable peptides. Therefore, to achieve data balancing I will continue to add more sequences to dataset.

This project aims to expand the scope of research that is carried out in the field of developing Antimicrobial peptides as therapeutics by making AMP prediction accessible and user friendly. While in the first half of the project I spent most of my time in analysing databases with little to detailed information on discovered AMPs. To commence with, I scrapped the data, cleansed it, analysed the sequence length, net charge affect, hydrophobic ratio impact, alignment of clean data using BLAST and CLUSTALW. The results were however insignificant and inconclusive due to the large amount of data. Before training the model it was essential to ensure there is no imbalance in data. To enhance prediction, I assessed all the evident parameters known to AMPs.

The inspiration of this project is drawn from the rapid and concerning cases of antimicrobial resistance against multiple antibiotics. In order to overcome this, the project will classify, predict and analyse other potential peptide sequences that can be developed clinically. Not only this, unlike other single label classification model my project work aims to achieve prediction through allowing the model to extract features assessing the amino acid percentage, hydrophobic amino acid ratio, net charge on peptides at isoelectric points and manually vectorising the amino acids to achieve binary relevance of features. The end goal is to create a tool that not only predicts antibacterial, antifungal, antibiofilm peptides but strikes out non-AMPs as well and effectuate accuracy for synthetic AMP production as well while looking for conserved regions as well.

Using sequence-based features, such as compositional and binary profiles, we created a number of prediction models using alignment-free machine learning and deep learning techniques. We looked at every potential feature in an effort to accurately predict the different ABP groups. Using basic amino acid composition, we found that the model performed better than the models, it pre trained itself to understand the sequences. During my research, by means of several comparisons with notable instances of the most advanced AMP predictors, we present AMP-iT, a tool that yields top-ranked predictions. Surprisingly, when tested with any user input peptide sequence the AMP-iT finder yields the most accurate predictions.

Chapter 6: Conclusion

Antibacterial peptides show great promise as the next generation of antibiotics to combat the difficult issue of bacterial drug resistance. Our study's three categorization models were developed using a variety of machine learning techniques. These techniques include k-nearest neighbour (kNN), random forest (RF), decision tree (DT), logistic regression (kNN), support vector machine (SVM). We utilised the well-known Python machine learning library, Scikit-learn, to create these classifiers. [34]

A thorough framework for AMP prediction has been made possible by the integration of various techniques, emphasising crucial elements like sequence order, charge, and hydrophobicity that are necessary for antimicrobial activity. This work provides important insights into the synthesis of novel AMPs and expands our knowledge of AMP functions, opening the door to creative treatment approaches to counteract infections that are resistant to drugs. In order to increase prediction accuracy and utility, future work will concentrate on improving current models and investigating new characteristics. Developing the AMP predictor tool with easy user interface and comprehensive peptide related stats and information will be integrated in a web application.

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT
PLAGIARISM VERIFICATION REPORT

Date: 7/06/24

Type of Document (Tick): PhD Thesis M.Tech/M.Sc. Dissertation B.Tech./B.Sc./BBA/Other

Name: DRISHTI AWASTHI Department: BIOTECHNOLOGY Enrolment No 201817

Contact No. 8077253831 E-mail. drishawa2503@gmail.com

Name of the Supervisor: DR. JITENDRA VASHIST

Title of the Thesis/Dissertation/Project Report/Paper (In Capital letters): PREDICTION OF ANTIMICROBIAL PEPTIDES USING MACHINE LEARNING APPROACHES

UNDERTAKING

I undertake that I am aware of the plagiarism related norms/ regulations, if I found guilty of any plagiarism and copyright violations in the above thesis/report even after award of degree, the University reserves the rights to withdraw/ revoke my degree/report. Kindly allow me to avail Plagiarism verification report for the document mentioned above.


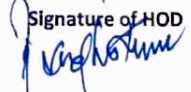
- Total No. of Pages = 48
- Total No. of Preliminary pages = 10
- Total No. of pages accommodate bibliography/references = 3


(Signature of Student)

FOR DEPARTMENT USE

We have checked the thesis/report as per norms and found Similarity Index at 10.....(%). Therefore, we are forwarding the complete thesis/report for final plagiarism check. The plagiarism verification report may be handed over to the candidate.


(Signature of Guide/Supervisor)


07/06/2024
Signature of HOD


FOR LRC USE

The above document was scanned for plagiarism check. The outcome of the same is reported below:

Copy Received on	Excluded	Similarity Index (%)	Abstract & Chapters Details	
<u>7/06/24</u>	<ul style="list-style-type: none"> • All Preliminary Pages • Bibliography/Images/Quotes • 14 Words String 	<u>10%</u>	Word Counts	<u>7738</u>
Report Generated on			Character Counts	<u>57696</u>
<u>7/06/24</u>		Submission ID	Page counts	<u>48</u>
		<u>2397566445</u>	File Size	<u>4.2 MB</u>

Checked by
Name & Signature

Librarian

Please send your complete Thesis/Report in (PDF) & DOC (Word File) through your Supervisor/Guide at plagcheck.juit@gmail.com

References

1. Gallo RL, Huttner KM. Antimicrobial peptides: an emerging concept in cutaneous biology. *J Invest Dermatol.* 1998;111(5):739–43.
2. Su, X., Xu, J., Yin, Y., Quan, X., & Zhang, H. (2019, December 1). *Antimicrobial peptide identification using multi-scale convolutional network.* *BMC Bioinformatics.* <https://doi.org/10.1186/s12859-019-3327-y>
3. Das, P. et al. Accelerated antimicrobial discovery via deep generative models and molecular dynamics simulations. *Nat. Biomed. Eng.* 5, 613–623 (2021).
4. Huan Y, Kong Q, Mou H, Yi H. Antimicrobial Peptides: Classification, Design, Application and Research Progress in Multiple Fields. *Front Microbiol.* 2020 Oct 16;11:582779. doi: 10.3389/fmicb.2020.582779. PMID: 33178164; PMCID: PMC7596191.
5. Huang, J., Xu, Y., Xue, Y., Huang, Y., Li, X., Chen, X., Xu, Y., Zhang, D., Zhang, P., Zhao, J., & Ji, J. (2023, January 12). *Identification of potent antimicrobial peptides via a machine-learning pipeline that mines the entire space of peptide sequences.* *Nature Biomedical Engineering.*
6. Wang G, Vaisman II, van Hoek ML. Machine Learning Prediction of Antimicrobial Peptides. *Methods Mol Biol.* 2022;2405:1-37. doi: 10.1007/978-1-0716-1855-4_1. PMID: 35298806; PMCID: PMC9126312.
7. Luo, X., Chen, H., Song, Y., Qin, Z., Xu, L., He, N., Tan, Y., & Dessie, W. (2023, February 1). *Advancements, challenges and future perspectives on peptide-based drugs: Focus on antimicrobial peptides.* *European Journal of Pharmaceutical Sciences.* <https://doi.org/10.1016/j.ejps.2022.106363>
8. Santos, C. D., R, R. G., F, L. L., MCG, D. R., NB, C., SC, D., & L, F. O. (2022, December 12). *Advances and perspectives for antimicrobial peptide and combinatorial therapies.* *Frontiers in Bioengineering and Biotechnology.* <https://doi.org/10.3389/fbioe.2022.1051456>

9. Duong, L., Gross, S. P., and Siryaporn, A. (2021). Developing antimicrobial synergy with AMPs. *Front. Med. Technol.* 3, 640981. doi:10.3389/fmedt.2021.640981
10. Kalita, A., Verma, I., and Khuller, G. K. (2004). Role of human neutrophil peptide-1 as a possible adjunct to antituberculosis chemotherapy. *J. Infect. Dis.* 190, 1476–1480. doi:10.1086/424463
11. Santos, C. D., R, R. G., F, L. L., MCG, D. R., NB, C., SC, D., & L, F. O. (2022, December 12). *Advances and perspectives for antimicrobial peptide and combinatory therapies*. *Frontiers in Bioengineering and Biotechnology*. <https://doi.org/10.3389/fbioe.2022.1051456>
12. Luong HX, Thanh TT, Tran TH. Antimicrobial peptides - Advances in development of therapeutic applications. *Life Sci.* 2020 Nov 1;260:118407. doi: 10.1016/j.lfs.2020.118407. Epub 2020 Sep 12. PMID: 32931796; PMCID: PMC7486823.
13. Rajagopal M, Walker S. Envelope structures of gram-positive bacteria. *Curr Top Microbiol Immunol.* 2017;404:1–44.
14. Baek MH, Kamiya M, Kushibiki T, Nakazumi T, Tomisawa S, Abe C, et al. Lipopolysaccharide-bound structure of the antimicrobial peptide cecropin P1 determined by nuclear magnetic resonance spectroscopy. *J Pept Sci.* 2016;22(4):214–21.
15. Malanovic N, Lohner K. Gram-positive bacterial cell envelopes: the impact on the activity of antimicrobial peptides. *Biochim Biophys Acta.* 2016;1858(5):936–46.
16. Lee TH, Hall KN, Aguilar MI. Antimicrobial peptide structure and mechanism of action: a focus on the role of membrane structure. *Curr Top Med Chem.* 2016;16(1):25–39.
17. Silva JP, Appelberg R, Gama FM. Antimicrobial peptides as novel anti-tuberculosis therapeutics. *Biotechnol Adv.* 2016;34(5):924–40.
18. Zhang, Q. Y., Yan, Z. B., Meng, Y. M., Hong, X. Y., Shao, G., Ma, J. J., Cheng, X. R., Liu, J., Kang, J., & Fu, C. Y. (2021, September 9). *Antimicrobial peptides:*

- mechanism of action, activity and clinical potential*. Military Medical Research/ Military Medical Research. <https://doi.org/10.1186/s40779-021-00343-2>
19. Dennison SR, Harris F, Mura M, Phoenix DA. An atlas of anionic antimicrobial peptides from amphibians. *Curr Protein Pept Sci*. 2018;19(8):823–38.
 20. Chen, Y., Guarnieri, M. T., Vasil, A. I., Vasil, M. L., Mant, C. T., & Hodges, R. S. (2007). Role of peptide hydrophobicity in the mechanism of action of alpha-helical antimicrobial peptides. *Antimicrobial agents and chemotherapy*, 51(4), 1398–1406. <https://doi.org/10.1128/AAC.00925-06>
 21. Wang, G., Li, X. and Wang, Z. (2016) APD3: the antimicrobial peptide database as a tool for research and education. *Nucleic Acids Research* 44, D1087-D1093.
 22. Gogoladze G., Grigolava M., Vishnepolsky B., Chubinidze M., Duroux P., Lefranc M.P., Pirtskhalava M.. DBAASP: Database of antimicrobial activity and structure of peptides. *FEMS Microbiol. Lett*. 2014; 357:63–68. - PubMed
 23. Wang G, Vaisman II, van Hoek ML. Machine Learning Prediction of Antimicrobial Peptides. *Methods Mol Biol*. 2022;2405:1-37. doi: 10.1007/978-1-0716-1855-4_1. PMID: 35298806; PMCID: PMC9126312
 24. Chia-Ru Chung, Jhen-Ting Liou, Li-Ching Wu, Jorng-Tzong Horng, and Tzong-Yi Lee, "Multi-label classification and features investigation of antimicrobial peptides with various functional classes," **iScience**, vol. 26, no. 12, p. 108250, 2023. [Online]. Available: <https://doi.org/10.1016/j.isci.2023.108250>.
 25. Mukhopadhyay, S., Bharath Prasad, A. S., Mehta, C. H., & Nayak, U. Y. (2020). Antimicrobial peptide polymers: no escape to ESKAPE pathogens-a review. *World journal of microbiology & biotechnology*, 36(9), 131. <https://doi.org/10.1007/s11274-020-02907-1>
 26. Huan, Y., Kong, Q., Mou, H., & Yi, H. (2020). Antimicrobial Peptides: Classification, Design, Application and Research Progress in Multiple Fields. *Frontiers in microbiology*, 11, 582779. <https://doi.org/10.3389/fmicb.2020.582779>

27. Lazzaro, B. P., Zasloff, M., & Rolff, J. (2020). Antimicrobial peptides: Application informed by evolution. *Science (New York, N.Y.)*, 368(6490), eaau5480. <https://doi.org/10.1126/science.aau5480>
28. Dietterich T.G. *Ensemble Methods in Machine Learning*. Springer; Berlin/Heidelberg, Germany: 2000. pp. 1–15.
29. García-Jacas CR, Pinacho-Castellanos SA, García-González LA, Brizuela CA. Do deep learning models make a difference in the identification of antimicrobial peptides? *Brief Bioinform*. 2022 May 13;23(3):bbac094. doi: 10.1093/bib/bbac094. PMID: 35380616.
30. Timmons, P. B., & Hewage, C. M. (2020). HAPPENN is a novel tool for hemolytic activity prediction for therapeutic peptides which employs neural networks. *Scientific reports*, 10(1), 10869. <https://doi.org/10.1038/s41598-020-67701-3>
31. Bhadra P, Yan J, Li J, Fong S, Siu SWI. AmPEP: Sequence-based prediction of antimicrobial peptides using distribution patterns of amino acid properties and random forest. *Sci Rep*. 2018 Jan 26;8(1):1697. doi: 10.1038/s41598-018-19752-w. PMID: 29374199; PMCID: PMC5785966.
32. M.-L. Zhang and Z.-H. Zhou, "ML-KNN: A lazy learning approach to multi-label learning," in *Pattern Recognition*, vol. 40, no. 7, pp. 2038-2048, 2007. <https://doi.org/10.1016/j.patcog.2006.12.019>.
33. Lee, H., Lee, S., Lee, I., & Nam, H. (2023). AMP-BERT: Prediction of antimicrobial peptide function based on a BERT model. *Protein science : a publication of the Protein Society*, 32(1), e4529. <https://doi.org/10.1002/pro.4529>
34. Bajiya, N., Choudhury, S., Dhall, A., & Raghava, G. P. S. (2024). AntiBP3: A Method for Predicting Antibacterial Peptides against Gram-Positive/Negative/Variable Bacteria. *Antibiotics (Basel, Switzerland)*, 13(2), 168. <https://doi.org/10.3390/antibiotics13020168>

35. Bogusz D, Appleby CA, Landsmann J, Dennis ES, Trinick MJ, Peacock WJ. Functioning haemoglobin genes in non-nodulating plants. *Nature*. 1988 Jan 14;331(6152):178-80. doi: 10.1038/331178a0. PMID: 2448639.
36. Monincová L, Budesínský M, Slaninová J, Hovorka O, Cvacka J, Voburka Z, Fucík V, Borovicková L, Bednářová L, Straka J, Cerovský V. Novel antimicrobial peptides from the venom of the eusocial bee *Halictus sexcinctus* (Hymenoptera: Halictidae) and their analogs. *Amino Acids*. 2010 Aug;39(3):763-75. doi: 10.1007/s00726-010-0519-1. Epub 2010 Mar 3. PMID: 20198492.
37. Ali MF, Soto A, Knoop FC, Conlon JM. Antimicrobial peptides isolated from skin secretions of the diploid frog, *Xenopus tropicalis* (Pipidae). *Biochim Biophys Acta*. 2001 Nov 26;1550(1):81-9. doi: 10.1016/s0167-4838(01)00272-2. PMID: 11738090.
38. Kerkis A, Kerkis I, Rádis-Baptista G, Oliveira EB, Vianna-Morgante AM, Pereira LV, Yamane T. Crotonamine is a novel cell-penetrating protein from the venom of rattlesnake *Crotalus durissus terrificus*. *FASEB J*. 2004 Sep;18(12):1407-9. doi: 10.1096/fj.03-1459fje. Epub 2004 Jul 1. PMID: 15231729.
39. Nascimento FD, Hayashi MA, Kerkis A, Oliveira V, Oliveira EB, Rádis-Baptista G, Nader HB, Yamane T, Tersariol IL, Kerkis I. Crotonamine mediates gene delivery into cells through the binding to heparan sulfate proteoglycans. *J Biol Chem*. 2007 Jul 20;282(29):21349-60. doi: 10.1074/jbc.M604876200. Epub 2007 May 9. PMID: 17491023.
40. Cammue BP, De Bolle MF, Terras FR, Proost P, Van Damme J, Rees SB, Vanderleyden J, Broekaert WF. Isolation and characterization of a novel class of plant antimicrobial peptides from *Mirabilis jalapa* L. seeds. *J Biol Chem*. 1992 Feb 5;267(4):2228-33. PMID: 1733929.
41. Mandard N, Sodano P, Labbe H, Bonmatin JM, Bulet P, Hetru C, Ptak M, Vovelle F. Solution structure of thanatin, a potent bactericidal and fungicidal insect peptide, determined from proton two-dimensional nuclear magnetic resonance data. *Eur J Biochem*. 1998 Sep 1;256(2):404-10. doi: 10.1046/j.1432-1327.1998.2560404.x. PMID: 9760181.

42. Dash, R., & Bhattacharjya, S. (2021). Thanatin: An Emerging Host Defense Antimicrobial Peptide with Multiple Modes of Action. *International journal of molecular sciences*, 22(4), 1522. <https://doi.org/10.3390/ijms22041522>
43. Khazaal NM, Alghetaa HF, Al-Shuhaib MBS, Al-Thuwaini TM, Alkhammas AH. A novel deleterious oxytocin variant is associated with the lower twinning ratio in Awassi ewes. *Anim Biotechnol*. 2023 Dec;34(8):3404-3415. doi: 10.1080/10495398.2022.2152038. Epub 2022 Nov 30. PMID: 36449364.
44. Sabri, M., El Handi, K., Valentini, F., De Stradis, A., Cara, O., Calvano, C. D., Bianco, M., Trani, A., & Elbeaino, T. (2024). Nisin-based therapy: a realistic and eco-friendly biocontrol strategy to contrast *Xylella fastidiosa* subsp. *pauca* infections in planta. *Frontiers in microbiology*, 15, 1406672. <https://doi.org/10.3389/fmicb.2024.1406672>