Fall 2020

Prediction of Novel Antibiofilm Peptides from Diverse Habitats using Machine Learning

Bipasa Bose
San Jose State University

PREDICTION OF NOVEL ANTIBIOFILM PEPTIDES FROM DIVERSE HABITATS USING MACHINE LEARNING

A Thesis

Presented to

The Faculty of the Department of Biomedical Engineering

San José State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Bipasa Bose

December 2020

The Designated Thesis Committee Approves the Thesis Titled

PREDICTION OF NOVEL ANTIBIOFILM PEPTIDES FROM DIVERSE HABITATS USING MACHINE LEARNING

by

Bipasa Bose

APPROVED FOR THE DEPARTMENT OF BIOMEDICAL ENGINEERING

SAN JOSÉ STATE UNIVERSITY

December 2020

Dr. Anand Ramasubramania, Ph.D.	Department of Chemical Engineering
Dr. David C. Anastasiu, Ph.D.	Department of Computer Science and Engineering, Santa Clara University
Dr. Folarin Erogbogbo, Ph.D.	Department of Biomedical Engineering

ABSTRACT

PREDICTION OF NOVEL ANTIBIOFILM PEPTIDES FROM DIVERSE HABITATS USING MACHINE LEARNING

by Bipasa Bose

Multidrug resistant bacteria often lead to biofilm formation. Biofilm is a colonized form of pathogens (fungi, bacteria) attached to surfaces like animal or plant tissues, medical devices like catheters, and artificial heart valves. Biofilm formation prolongs the survival of microorganisms in an adaptive environment, leading to the spread of infection in different organs and causing a high morbidity rate. Given the rise of chronic infection and antibiotic resistance due to biofilm, it is essential to find an alternative solution to control biofilm infections. Antibiofilm peptides can interact with these biofilm-creating pathogens to inhibit growth, virulence, and biofilm formation. We hypothesized that mining the existing peptide databases from diverse habitats could provide potential antibiofilm activities for our work. We developed a computational model to predict the antibiofilm properties by applying machine learning algorithms like support vector machine, random forest, extreme gradient boosting, and multilayer perceptron classifier. We evaluated more than 240 antibiofilm peptides and more than 570 different compositions and motif-based features to build our prediction model. We also created a regression model on top of our classifier to predict the effectiveness of peptides by curating minimum inhibitory concentration against biofilm. Our classifiers achieved greater than 98% accuracy while the harmonic mean of precision-recall (F1) and Matthews correlation coefficient (MCC) scores obtained are greater than 0.91. Using this two-tier model approach, we assessed more extensive databases of antimicrobial, anticancer, antiviral, and dairy peptides for potential antibiofilm functionality and came up with the top ten potential candidates of antibiofilm peptides.

ACKNOWLEDGMENTS

I would like to thank my advisors Dr. Anand Ramasubramania and Dr. David C. Anastasiu, for their support and guidance throughout my research. This work would not have been possible without their valuable inputs. I would also like to thank my reading committee, Dr. Folarin Erogbogbo, for his time and suggestions.

My sincere thanks to Cheryl Cowan, Graduate Studies Associate, Dr. Jeffrey Honda, Associate Dean of Graduate Programs, and Dr. Marc d'Alarcao, Dean of the College of Graduate Studies, for granting the extension. I do appreciate Cheryl's workshop on thesis formatting, which helped greatly in the formatting of this work.

This work would not have been possible without the support of my family.

DEDICATION

I would like to dedicate this work to my beloved family.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AMP	Antimicrobial Peptides
APD	Antimicrobial Peptide Database
BaAMP	Biofilm Active Peptide Database
DRAMP	Data Repository of Antimicrobial Peptides
MBIC	Minimum Biofilm Inhibitory Concentration
MDR	Multidrug Resistance
MIC	Minimum Inhibitory Concentration
MLP	Multilayer Perceptron
RF	Random Forest
SVM	Support Vector Machines
XGBoost	Extreme Gradient Boosting

1 INTRODUCTION

Multidrug resistant (MDR) infection sickens millions of people each year. According to a recent report published by the Centers for Disease Control and Prevention (CDC), each year, approximately 3 million people face infection due to antibiotic resistance in the U.S., and around 35,000 people die as a result [1]. MDR is the antimicrobial resistance of microorganisms against treatment with multiple antibacterial drugs. According to the World Health Organization (WHO,) microorganisms (bacteria, fungi, yeast) manage to remain insensitive mainly due to genetic changes with a broad set of current antibiotics, including third-generation drug like cephalosporins. Microorganisms (especially gram-negative bacteria) grow in-built power to efflux out the drugs, leading to cell survival. Also, they can pass genetic material to other bacteria to build the same immunity. This problem leads the infection to be untreated and spread from one organ to another, severely damaging tissues. These infecting microorganisms develop a high resistance level with increased morbidity and mortality rate, sometimes termed ”superbug” [2]. The European Food Safety Authority spent almost 1.5 billion euro each year due to antibiotic resistant infections. WHO also has projected multidrug resistance imposes a substantial financial crisis globally [3]. There has been extensive research to find an alternative solution against these ”superbugs” which has lead to finding antimicrobial peptides (AMPs).

AMPs are one of the significant components of the innate immune defense systems [4] found in bacteria, fungi, plants, and animals. Studies show that antibacterial activity is one of the main features of AMPs. Bacteria do not generally develop resistance to this type of peptide as these peptides can physically disrupt bacterial growth by killing them. The amphipathic nature of AMPs makes them an outstanding candidate for dealing with antidrug resistance. Our current work focuses on a subset of AMPs called the antibiofilm peptide.

Multidrug resistant bacteria often lead to the formation of biofilm. Biofilm formation prolongs the survival of microorganisms in an adaptive environmental niche [5]. Biofilm is a colonized form of pathogens (fungi or bacteria) formed on surfaces like household surfaces, contact lenses, or medical devices like catheters and artificial valves. Biofilm is one of the main reasons for antibiotic resistance in most intensive care units of hospitals. *Pseudomonas aeruginosa* biofilm is one of the main reasons for a chronic lung infection in cystic fibrosis patients [6]. The mucoid biofilm shows resistance against antibiotics and other innate immune defense systems, causing prolonged inflammation and damage to the lung tissues. Biofilm formation on medical device surfaces can be life-threatening and often lead to device failure and chronic infection. Urinary catheters (25-35%) and dental implants (10-56%) are most prone to bacterial colonization [7].

Biofilm growth can be divided into various stages. Fig. 1 depicts different stages of biofilm life cycle (adapted from [8]). In the first stage, a moving bacterial (planktonic) cell comes in contact with the surface. In the next phase, the cell starts to grow and build a colony called a microcolony. The microcolony develops a slime like environment called extracellular polymeric substances (EPS) containing polysaccharides. The extracellular substance supports the growth and nutrient transfer to other bacteria. When the colony development reaches multiple layers of formation, the formation becomes irreversible. The irreversible growth of the colony and EPS turns into a 3D shaped territory. The 3D 'mushroom' shaped colony consists of microbes, EPS, ions, enzymes, proteins, and nucleic acids. In the mature stage, one of the colony members detaches itself and disperses as a single (planktonic) cell to start another colony cycle.

When the biofilm reaches its mature phase, it becomes resistant to most antibiotics and is hard to treat. Many organisms produce peptides in their natural environment that interact with these microorganisms to inhibit growth, virulence, and biofilm formation. These peptides are called antibiofilm peptides. An antibiofilm peptide is effective against

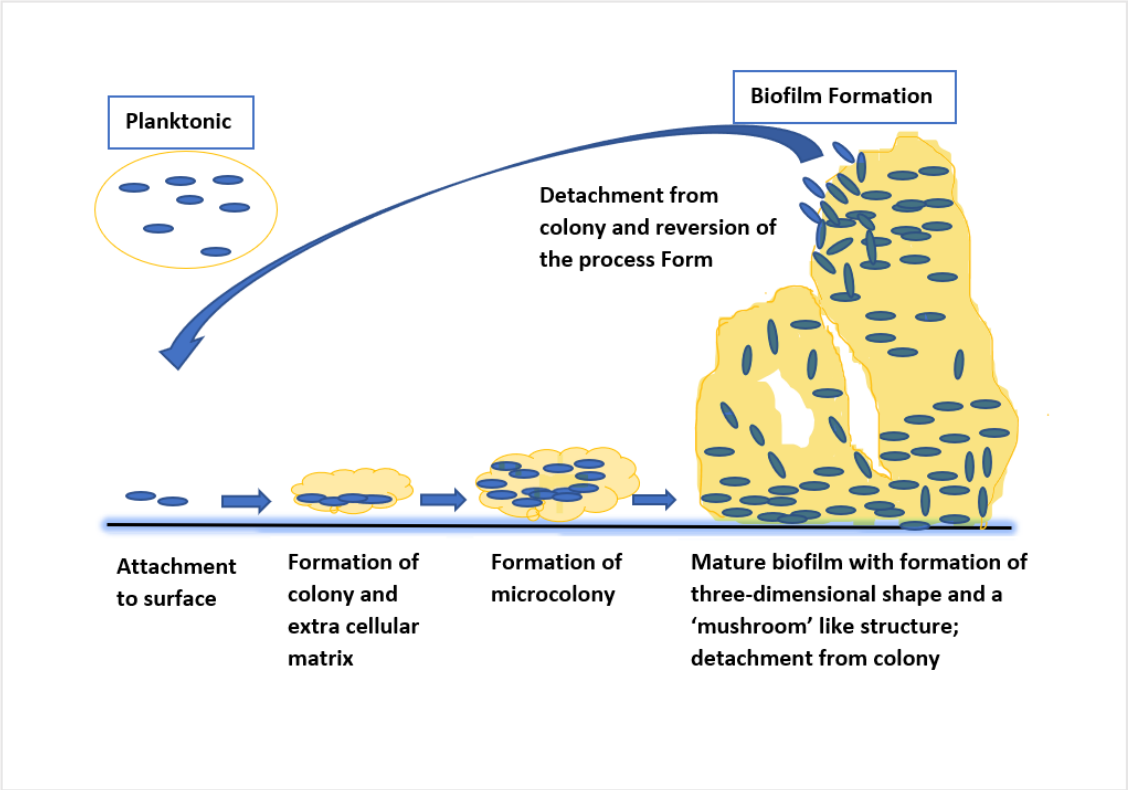

Fig. 1. Different phases of biofilm formation.

biofilm when the minimum biofilm inhibitory concentration (MBIC) is lower than the minimum inhibition concentration (MIC) against planktonic bacteria [9]. The antibiofilm peptides also possess a similar bacterial killing capacity as AMPs. Antibiofilm peptides could be significant in two ways. Antibiofilm peptides can inhibit biofilm growth to the attached surface as well as disrupt biofilm formation by eradicating the pathogen. The effectiveness of antibiofilm peptides against preforming biofilm is measured with minimum biofilm eradication concentration (MBEC).

Biofilm is one of the primary sources of chronic and deadly infections. Due to its resistance against most antibiotics, the current treatment option involves a combination of more than one antibiotic with a higher dose, increasing the risk of cytotoxicity [10]. The

real challenge in the fight against bacterial resistance is to develop a potential antibiofilm peptide by limiting cytotoxicity and increasing effectiveness. While working with a vast library of peptides is time-consuming and highly expensive, computational models bring hope to screen more novel antibiofilm peptides in a timely and cost-effective manner. The *in silico* approach to screen novel peptides became quite popular in the last decade with the increasing application of machine learning and artificial intelligence in the healthcare system.

In this work, we developed a computational model to detect peptides that have the potential power to inhibit biofilm growth. We hypothesized that mining various peptide databases for putative antibiofilm activity will provide peptides with diverse activity against human pathogens. There are very few data sources currently available to find peptides effective against biofilm. Databases like Biofilm Active Peptide Database (BaAMPs) [11] and Antimicrobial Peptide Database (APD) [12] were used for our data collection purposes. We assessed various compositional features like 'amino acid composition,' 'dipeptides composition,' of antibiofilm peptides. We also screened various physical and chemical properties like charge, hydrophobicity, secondary structure, motif, and other features. We applied these sequence based and property based features of peptides to build our classification model. We applied different machine learning algorithms such as support vector machine (SVM), random forest (RF), extreme gradient boosting (XGBoost), and multilayer perceptron (MLP) classifier to select the best performance model of all. We also curated MBIC values for different antibiofilm peptides to build a regression model to predict peptide efficacy. Our final model for screening potential antibiofilm peptides was based on our classifier model, regression model, and motif analysis. The flow diagram of our prediction decision is illustrated in Fig. 2.

Though few existing computational models are built for screening antibiofilm peptides [13]–[15], we intended to improve the model by introducing new parameters and

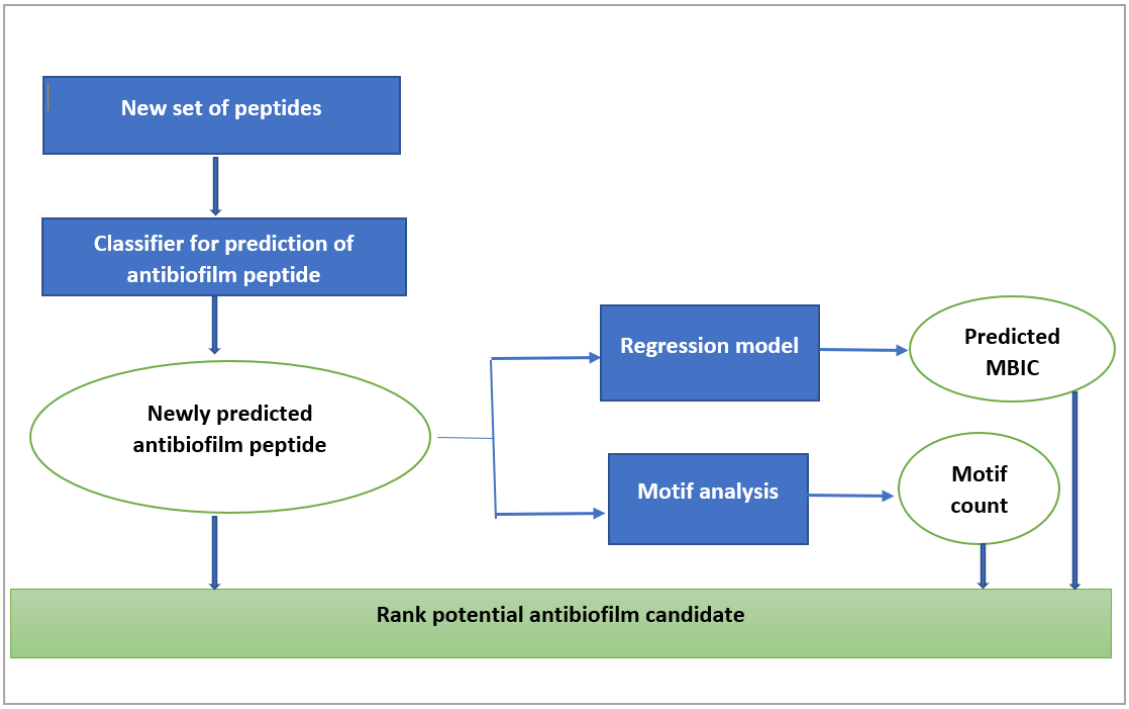

Fig. 2. The flow diagram to determine antibiofilm peptide using our two-tier computational model.

approaches. To date, we are not aware of any prediction model that considers peptides' effectiveness. Our new approach not only classifies a peptide as antibiofilm but also predicts the efficacy of it. So we believe this new approach will help screen thousands of peptides from different habitats and predict a novel candidate for the antibiofilm property.

2 LITERATURE REVIEW

2.1 Background

All living organisms like bacteria, fungi, plants, and animals are the source of antimicrobial peptides (AMPs). AMPs are the multifunctional component which is the first line of defense while fighting against the foreign environment, often called host defense peptides (HDPs) [16]. Antibiofilm peptides are a smaller subset of these HDPs. Lysozyme, discovered in 1922 by Sir Alexander Fleming, is considered the first AMP in history. Human cathelicidin LL-37 is the first peptide identified as antibiofilm [17]. To date, around 3000 AMPs have been discovered [12], while only a small fraction of them are considered antibiofilm. The source for antibiofilm peptides can be eukaryotic cells of fungi, plants, and animals. Synthetically derived peptides are also an excellent source for antibiofilm activity. The various origins of antibiofilm peptides from BaAMP database [11] are shown in Fig. 3.

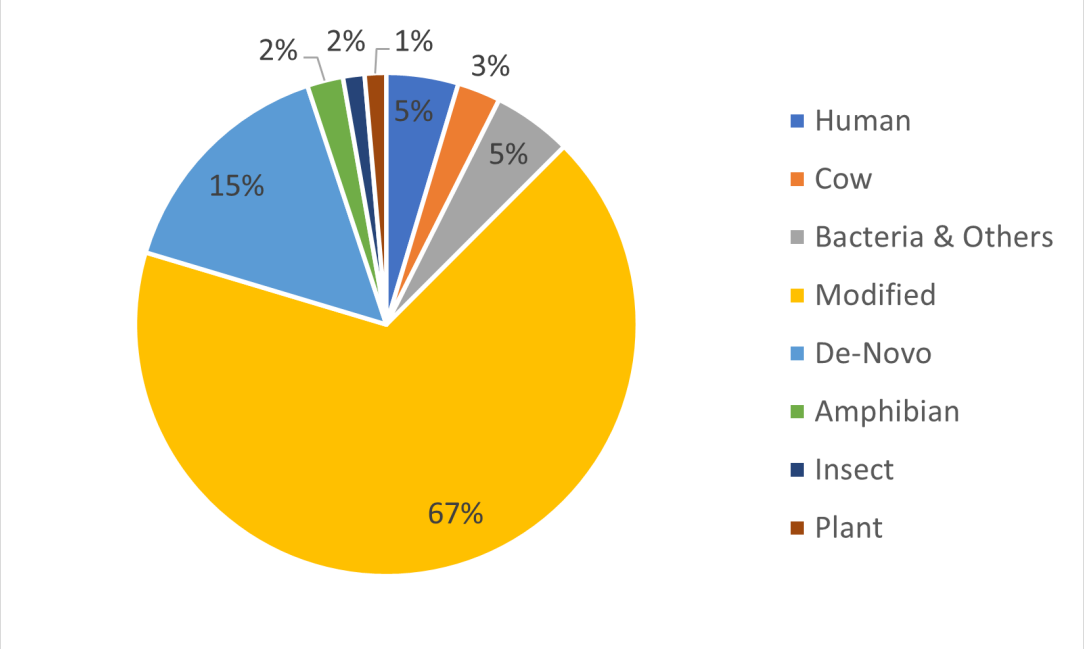

Fig. 3. Origins of antibiofilm peptides based on the BaAmp database.

Though AMPs and antibiofilm peptides vary in their origin, they share a few common characteristics:

- a highly positive net charge ranging between +9 to +12 due to increased presence of the amino acids like lysine and arginine [18]
- more than 50% hydrophobicity residue [19]
- a standard sequence length
- the ability to fold in amphipathic structure and targeting bacteria by membrane disruption

The mechanism of inhibition for antibiofilm peptides is yet not clear. Some studies show that antibiofilm peptides disrupt the cytoplasmic membrane associated function of microbes. With their catatonic residue, antibiofilm peptides contact the anionic phospholipid headgroup and insert it into the bacterial membrane. This cytoplasmic disruption activity is modeled as different structures like aggregate, barrel-stave and carpet [20]. It is also observed that the aromatic side chain of antibiofilm peptides helps to increase the concentration of the peptide molecules on the bacterial cell membrane and eventually detach the biofilm from the surface [21].

Antibiofilm peptide Nisin-A shows membrane disrupting behavior against *Staphylococcus aureus* (MRSA) biofilm [22]. The studies on human cathelicidin LL-37 show that it inhibits bacterial growth by disrupting the signaling system of *Pseudomonas aeruginosa*. LL-37 can reduce the initial attachment of bacterial cells to the surface and affect two quorum-sensing systems of bacteria: the Las and the Rhl [23]. The studies on synthetic peptides like 1018 show inhibition and eradication of the biofilm by modifying the stress response and degrading the biofilm forming signaling molecules like ppGpp [24]. Literature shows that the antibiofilm peptides act better against the gram-positive bacteria *Staphylococcus aureus* [25] compared to gram-negative ones.

2.2 Peptide Databases

There are few databases which are used as the source of antimicrobial peptide libraries. The databases are mainly manually curated and evolved with time. The APD has approximately 3175 AMP spread over 6 kingdoms (bacteria - 354, fungi - 20, plants - 352, animals - 2352 , archaea - 5, protists - 8 and a few synthetic ones). BaAMPs database contains information on AMPs that are specifically active against microbial biofilms. The database contains more than 200 AMPs. Another open-source library is DRAMP (Data Repository of Antimicrobial Peptides). This database includes AMPs from clinical trials and patents in their list [26]. The other database where we can find the quorum sensing peptides is Quorum Sensing Peptide Prediction Server [27]. The quorum sensing peptides are effective in biofilm formation. Some milk peptides have well known antimicrobial properties. Different antimicrobial peptides from the diary origin are listed in the MilkAMP database [28].

2.3 Existing Computational Models

The diverse functionality of antimicrobial peptides and their innate and host defense systems make them a great candidate for further research to fight against antibiotic resistance. Machine learning can predict data based on its model data set. The machine learning model exists for a long time. However, with modern computational advancement, research in machine learning exploded with some great algorithms like hidden markov models (HMM), support vector machines (SVM), and random forests (RF) [29]. Initially, the QSAR (quantitative structure active relationship) model was applied to AMPs to find the sequential peptides models [30]. Though this model successfully computed many physicochemical properties, it mainly depended on statistical learning, which made its scope smaller with other computational algorithms' advancement.

In 2011, P. Wang et al. worked on feature selection and the sequence alignment model to predict novel AMPs [31]. They achieved the Mathew's correlation coefficient (MCC)

value up to 0.73. In 2013, Xiao et al. came up with two-step machine-learning algorithms using K-nearest neighbor to predict antimicrobial peptides. Their novel classifier was named iAMP-2L. They also built a web server for user convenience on top of their classifier [32].

In 2017, Schneider et al. came up with the first-ever concept of deep neural networks to build a model for both 'unsupervised–supervised' learning. They tested their algorithm by considering the chemical properties of all alpha-helix AMPs from the APD database. They validated the functional prediction of their model experimentally [33]. In 2018, a novel 3D model based on AMP's structure was developed by Liu et al. to predict the activity and design of AMPs. Using their novel model, they designed 5 AMPs, and their antibacterial activity was verified experimentally in the lab using the *in-vitro* method [34].

Compared to work on antimicrobial peptides, not much computational model has been developed on antibiofilm peptides. In 2016, Gupta et al. came up with a sequence-based feature model for biofilm inhibiting peptides based on data from the BaAMP database. They choose the SVM and RF-based models to predict the antibiofilm activity. The SVM model proved to be most effective for them with a combination of DPC and motif features. They built a web server called Biofin on top of their best model, which achieved an MCC score of 0.84 [13].

Sharma et al. came up with another web server-based approach named DPABBs. The DPABBs web server was built on 6 SVM and Weka-based models and achieved an MCC of 0.91 [14]. They worked with BaAMP data for their positive dataset and quorum-sensing peptides as their negative set.

Recently, another web-based model, BIPEP was developed by Fallah Atanaki et al. They used a combination of peptide datasets from APD and BaAMP databases. Their negative dataset was smaller than the positive set and based on quorum sensing peptides. The model achieved an MCC value of 0.89 while using the SVM machine learning

algorithm. They considered a larger positive dataset than the negative dataset [15] and validated their model against an independent dataset.

All the above studies show great potential in machine learning prediction algorithms, which have been used with different approaches so far. The current models can be further enhanced to design and predict novel peptides. As we saw from the literature, we also considered machine learning algorithms like SVM, random forest in our work and used different databases like BaAMP, APD, etc., for our model.

3 METHODS & MATERIALS

3.1 Dataset Preparation

For our work, we collected datasets from different databases. The details of the dataset collection and data processing are described in the section below.

3.1.1 Main Dataset

For this work, we extracted antibiofilm peptides from the APD and BaAMP databases. Both databases are available for open access. After removing the duplicates, we had a dataset of 242 antibiofilm peptides as our positive set. The sequence length for the positive dataset varies between 4 and 69. For the negative dataset, we generated random peptides between a sequence length of 4 and 70. To make a realistic model, we used ten times more peptides in our negative dataset. The whole dataset, 2662 (242+2420) peptides, were used for training and testing with an 80:20 ratio. Eighty percent of the data was used for training and ten-fold cross-validation while the other 20% of the data was kept aside as an unknown set. We will refer to this set as our validation dataset for the rest of this thesis. The performances of different machine learning algorithms were evaluated on this validation dataset. The work with the main dataset will be referred to as our model A approach for the rest of this thesis.

3.1.2 Additional Dataset

We also compared our positive data against actual peptides, which are not antibiofilm. We analyzed the proteomic of different biofilm-forming bacteria like *Staphylococcus aureus* and *Escherichia coli* and found out a few biofilm-causing peptides from the NCBI and UniProt databases. For example, we considered Fibronectin-binding protein B, which promotes the accumulation and surface attachment of biofilm by *Staphylococcus aureus*. We curated proteins like surface protein G of bacteria like *Staphylococcus aureus* and *Staphylococcus oralis* in our negative dataset. We also considered that the quorum sensing

peptides act as biofilm forming agents. We used peptides from QSPProd in our negative dataset as well. QSPProd has already been used in literature [15]. We manually curated 1,100 such peptides to compare the performance of our model against naturally occurring peptides. We randomly generated the rest of the peptides to achieve ten times more peptides than antibiofilm peptides in the negative dataset. The additional dataset was a combination of manually curated and randomly generated peptides. The positive and the negative (additional) datasets were again distributed in 80:20 ratio for training and testing. The additional dataset will be referred to as model B approach for the rest of this thesis.

3.1.3 Dataset 2

We manually curated the minimum biofilm inhibitory concentration (MBIC) value for the antibiofilm peptides from our positive dataset. We curated 160 such values from literature against different gram-positive and gram-negative bacteria. We did not consider the rest of the peptides in our dataset due to their lower activity against biofilm and higher MBIC value. We believe the MBIC value is a significant indication of biofilm inhibition. Due to the lack of data on biofilm eradication concentration, we could not consider that parameter in our scope of work.

3.1.4 Test Data

The test datasets were collected from various sources. We collected 74 anticancer peptides, 220 antiviral peptides, and more than 4770 antimicrobial peptides from the DRAMP (Data Repository of Antimicrobial Peptides) database. We also considered 200 peptides from the MilkAMP database as our test dataset.

3.2 Feature Extraction

While working on a computational model, one of the primary goals is to analyze different peptide features like composition and structure. Peptide sequence based descriptors are essential parameters that have been in use for many machine learning

experiments [35]. Feature extraction is an essential part of developing any machine learning model. Feature extraction is the way of converting peptide information with numerical values. We used a few existing packages like 'propy3' [36], 'protParam' [37] to extract different features.

3.2.1 Amino Acid Composition (AAC)

The amino acid composition represents the fraction of each amino acid present in the peptide sequence. There are twenty known amino acids present in nature. The peptides of our interest are made of different combinations of these twenty amino acids. The AAC feature helps to calculate the percentage of each type of amino acid present in peptides. The python package returns a size twenty vector of named amino acids. The below equation represents the amino acid composition function:

$$AAC(i) = \frac{Total\ number\ of\ amino\ acid\ of\ type\ (i)}{Total\ number\ of\ amino\ acids} * 100 \tag{1}$$

3.2.2 Dipeptide Composition (DPC)

Dipeptide composition represents the total number of dipeptides present in the peptide sequence. There are twenty standard amino acids, so the number of possible dipeptides could be up to 400 (20 * 20). The DPC feature returns 400 named vectors with a non zero value to those combinations where dipeptides are present. The DPC feature has been extracted from the 'PyPro' package and can be expressed as below equation :

$$DPC(i,j) = \frac{Total\ number\ of\ dipeptides\ of\ amino\ acid\ type\ (i\ and\ j)}{Total\ number\ of\ possible\ dipeptides\ available} * 100 \tag{2}$$

3.2.3 Composition/ Transition/ Distribution(CTD)

The CTD descriptor returns different physio-chemical properties of the peptides. The properties of peptides returned by this feature are: 'Hydrophobicity,' 'Normalized van der Waals Volume,' 'Polarity,' 'Polarizability,' 'Charge,' 'Secondary Structure' and 'Solvent Accessibility.' The possible 20 amino acids are divided into three groups (group - 1, 2 & 3) depending on their property and functionality. The Composition feature represents the percentage of each group of amino acids in the peptide sequence. The Transition feature represents the percentage of frequencies in which group 2 amino acids follow group1 amino acids. Similarly, this feature also represents the frequencies of group 2 amino acids, followed by group 3. The Distribution represents the percentage residue of each attribute present in the peptide in their first or 0th position, 25% residues, 50% residues, 75% residues, and 100% residues. The details of AAC, DPC and CTD descriptors are listed in Table 1.

Table 1
Descriptions of Different Feature Descriptors

Descriptor	**No. Of Features**	**Feature Type**
Amino Acid Composition (AAC)	20	Percentage of amino acids
Dipeptide Composition (DPC)	400	Percentage of dipeptides
Composition/ Transition/ Distribution (CTD)	147	Distribution and variation of physicochemical properties

3.2.4 Motif Feature

Motifs are the smaller amino acid sequence present in peptide or protein structure, which may represent a unique biological or chemical function. We used MERCI

software [38] to find the motif of antibiofilm peptides. MERCI software provides two scripts to extract motifs. One script can essentially find all the motifs that are present in the positive dataset and absent in the negative dataset. The script also helps to store the motifs in a specific format file. We used that script to discover and store motifs in our training dataset. We used the second script to find the motifs of the test dataset from already stored motifs. We collected all the default motifs for our datasets. The number of motifs found in peptides was considered as a motif feature. Motif analysis represents distinct patterns in antibiofilm peptides that are not present in non-antibiofilm peptides.

3.2.5 Other Features

We extracted some critical features using the 'ProtParam' [37] module. Using the 'ProteinAnalysis' module of 'ProtParam,' we pulled features like sequence length, molecular weight, aromaticity, isoelectric point. We evaluated these features to build our model due to the presence of higher cationic and aromatic amino acids in antibiofilm peptides.

3.3 Machine Learning Models

We developed our prediction model using machine learning algorithms like support vector machines (SVM), random forest (RF), extreme gradient boosting (XGBoost), and multilayer perceptron (MLP) classifier. We used different classifier algorithms to compare their relative performance. Our goal was to select the best designs with the highest performance against our dataset. We used the "Scikit-learn" [39] package to develop models for our work.

3.3.1 Support Vector Machine (SVM)

Support Vector Machine is a well-known classifier for supervised machine learning [40]. SVM is one of the most commonly used classifiers for peptide prediction. SVM works particularly well in binary classification. SVM offers very high accuracy

compared to other classifiers, such as logistic regression and decision trees. The data fed into SVM separates using a hyperplane, which can be linear to high dimensional space. SVM aims to orient the hyperplane so that the hyperplane could maintain the most distance (margin) from the closest members of both classes. Since our dataset is not huge, it is better to use a nonparametric method that SVM supports. SVM also supports nonlinear kernel structure for hyperparameter tuning of the model. We used a radial basis function (RBF) kernel for our model due to the nonlinear characteristic of the dataset. SVM is a robust model that can be used for both classification and regression. Our work also involved the application of SVM both as a classification model and regression model. We used SVM to classify a peptide as antibiofilm and support vector regression (SVR) to predict minimum biofilm inhibitory concentration (MBIC). Literature shows that SVM has performed exceptionally well in peptide prediction. Gupta et al. used this algorithm for their peptide prediction resulting in 97% accuracy.

3.3.2 Random Forest (RF)

The Random Forest model is an ensemble model of supervised machine learning consisting of many different decision trees. Each tree predicts its own decision to a class. Finally, using the voting method, the class with the most votes comes as a prediction result. The single decision trees are independent of each other. Each decision tree inside the model is constructed using bootstrap sampling. In each split, the Gini impurity index helps to decide the number of splits. This model also supports both regression and classification. RF has already been used to predict peptides and to solve other biological problems [41]. For an imbalanced dataset, RF may not be the best choice as a classifier, but we used this algorithm to compare the performance with other classification algorithms.

3.3.3 Extreme Gradient Boosting (XGBoost)

Extreme gradient boosting is comparatively a new model used in machine learning. This method implements gradient boosted decision trees for a lower execution speed and higher performance of the model. This ensemble model incorporates a gradient boosting approach where a new model will be created in each iteration to minimize the prior layer's error. XGBoost uses DMatrices, which can contain both the features and target. The gradient boosting algorithm also can be used for classification and regression problems. In our work also we used XGBClassifier and XGBRegressor for classification and prediction. XGBoost algorithm has regularization parameters that can be tuned to reduce the overfitting issue for an imbalanced dataset. This algorithm is also used in prior work for the prediction of peptides with an accuracy greater than 98% [42].

3.3.4 Multilayer Perceptron (MLP)

The Multilayer perceptron is a type of feed-forward artificial neural network. This classifier consists of an input, an output, and one or more hidden layers that can be tuned to improve the accuracy. The hidden layer depicts the number of neurons in the neural network. The neurons have weighted parameters (generally between 0 to 0.3) to tune the input signal. The layer after the input signal is called the hidden layer which is built on multiple neurons. The activation function that is used to generate the output signal is mostly nonlinear. This classifier is based on supervised learning and uses the backpropagation of data for training. MLP can be used for both classification and regression problems. This algorithm is particularly useful when the dataset is nonlinear. Due to the character of our dataset, we tried this algorithm for our classifier model. This algorithm is also known as the 'vanilla' neural network.

3.4 Cross-Validation and Stratified Sampling

We did the cross-validation of our training dataset using ten-fold cross-validation, where the entire dataset was divided into ten parts. One part was used for testing, and the

other nine parts were used for training. Then the process was iterated over ten times to achieve ten-fold validation. This ten-fold validation was used to address the over-fitting problem. Our dataset has ten negative peptides for each positive peptide; hence it is an imbalanced dataset. We used stratified sampling to ensure that each fold receives an equal percentage of positive and negative datasets while doing cross-validation. Stratified sampling guarantees that each subgroup within the population gets a proper representation of each sample. There is no way to ensure that each fold of cross-validation receives an equal percentage of data in random sampling. Consequently, a few folds can sometimes end up as low as one positive data and represent an inaccurate performance in random sampling. The stratified sampling ensured that the validation dataset had precisely 20% of positive data, i.e., 48 peptides, and 20% of negative peptides, i.e., 485 peptides. The stratified sampling ensured a 9:1 ratio between training and testing set in each fold of ten-fold cross-validation. Each fold had a balance of 174 peptides for training and 19 peptides for testing (9:1) for the positive dataset, and the same distribution had been applied to the negative dataset.

3.5 Performance Evaluation

To assess the performance of our model, we used ten-fold cross-validation with our training data. The cross-validation was performed using the stratified folds, as reported in the above section. We evaluated the performance of different machine learning techniques against our dataset using several statistical parameters and metrics like sensitivity (Sen), specificity (Spec), accuracy (Acc), Matthew's correlation coefficient (MCC), and harmonic mean of the precision-recall (F1) Score. The methods to calculate different performance matrics are listed in the below equations.

$$Specificity = \frac{TN}{FP + TN} \tag{3}$$

$$Sensitivity = \frac{TP}{TP \ + \ FN} \tag{4}$$

$$Accuracy = \frac{TP \ + \ TN}{TP \ + \ FP \ + \ TN \ + \ FN} \tag{5}$$

$$F1Score = \frac{TP}{TP \ + \ \frac{FP \ + \ FN}{2}} \tag{6}$$

$$MCC = \frac{(TP)(TN) \ - \ (FP)(FN)}{\sqrt{(TP+FP) \ (TP+FN) \ (TN+FP) \ (TN+FN)}} \tag{7}$$

Here, TP = True Positive, TN = True Negative, FP = False Positive, FN = False Negative.

4 RESULTS

We built our classifier model with two different datasets. Model A was built on 242 antibiofilm peptides (positive dataset) with randomly generated negative dataset. We curated an additional dataset where we combined 1,100 non-antibiofilm peptides, and 1,320 randomly generated peptides to make our dataset a realistic representation of peptides in nature. In the model B approach, we compared the 242 antibiofilm peptides against our additional dataset. We also built a regression model to predict the minimum biofilm inhibitory concentration of peptides on top of our classifier model. For the second tier of our model, we used only antibiofilm peptides with MBIC value lower or equal with 64 μM. We curated 160 such peptides and trained our regression model. Considering the prediction probability from the classifier, prediction from the regression model, and the motif count, we decided on potential new antibiofilm candidates. Fig. 4 depicts our model A flow diagram.

The flow diagram of model B is also described in Fig. 5. The difference between these two models is the data collection step. The other methodologies remain the same as model A.

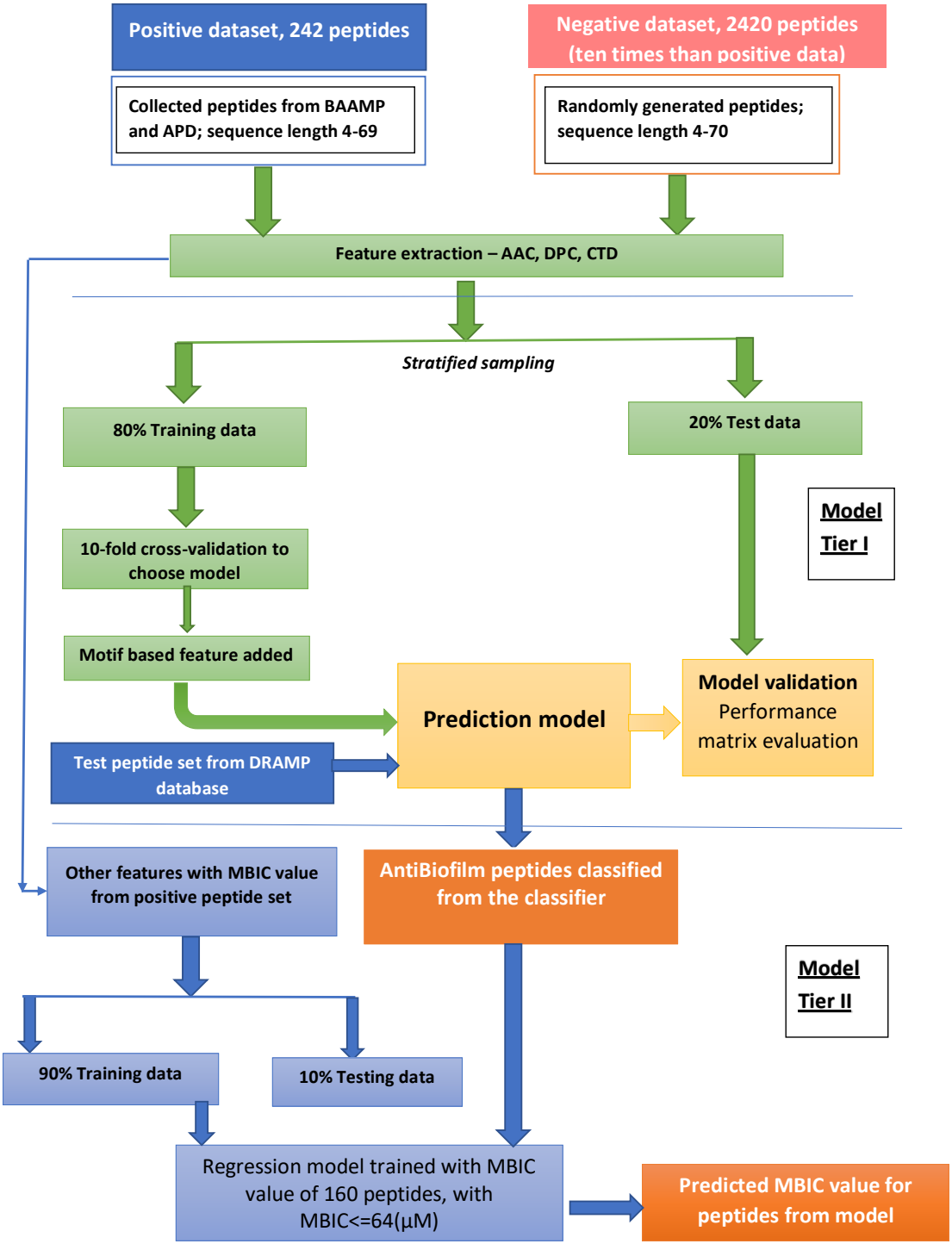

Fig. 4. Flowchart diagram for different steps of Model A.

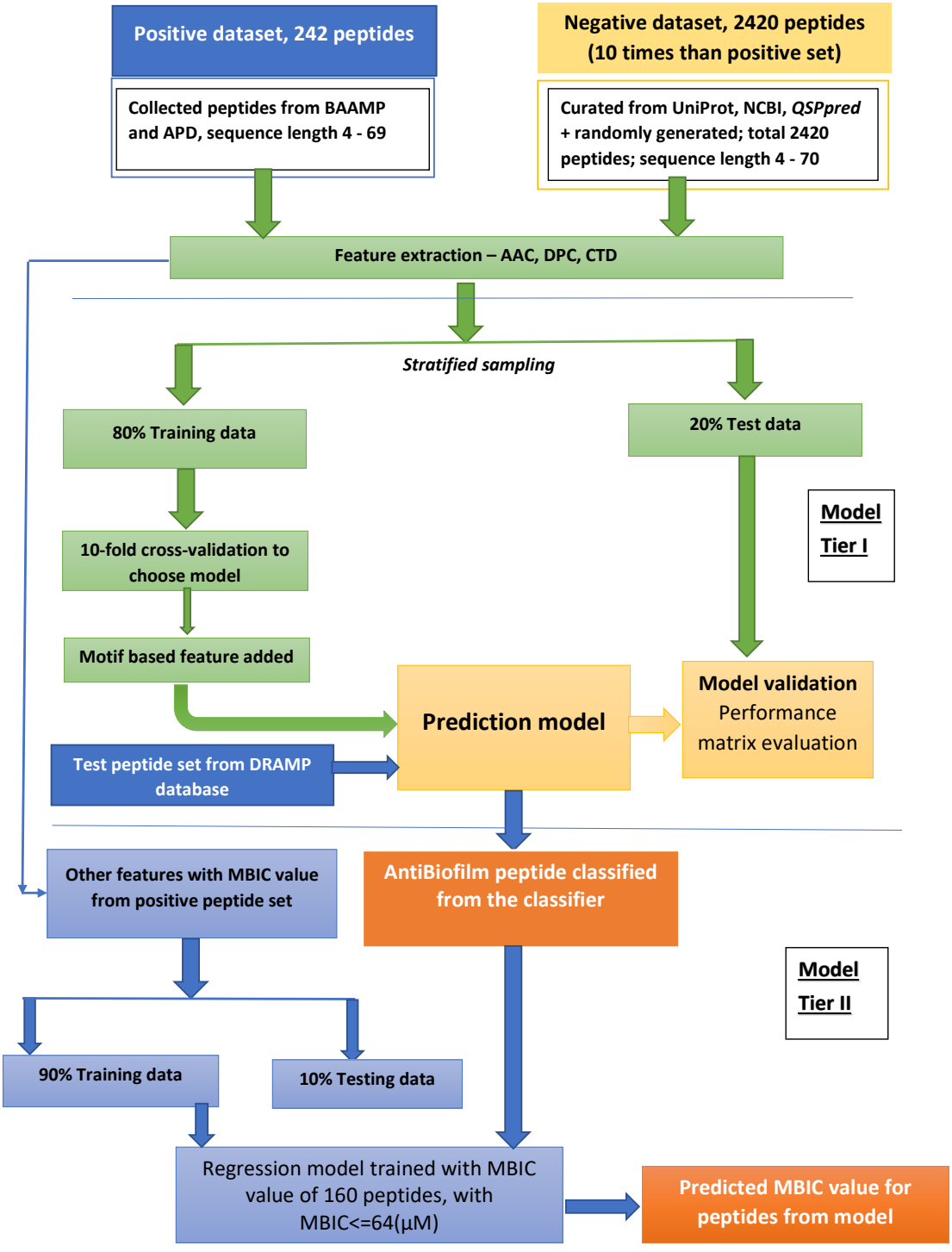

Fig. 5. Flowchart diagram for different steps of Model B.

4.1 Characteristics Of Positive Dataset

4.1.1 Sequence Length

The sequence length of different amino acids in our positive dataset is plotted in Fig. 6. The graph shows that most of the peptides have a sequence length between 10 and 16. Almost 43% of positive data belong in this group. Almost all the peptides have a sequence length less than 50. Only 2 peptides have a sequence length between 50-60 and 2 peptides have a sequence length between 60-70.

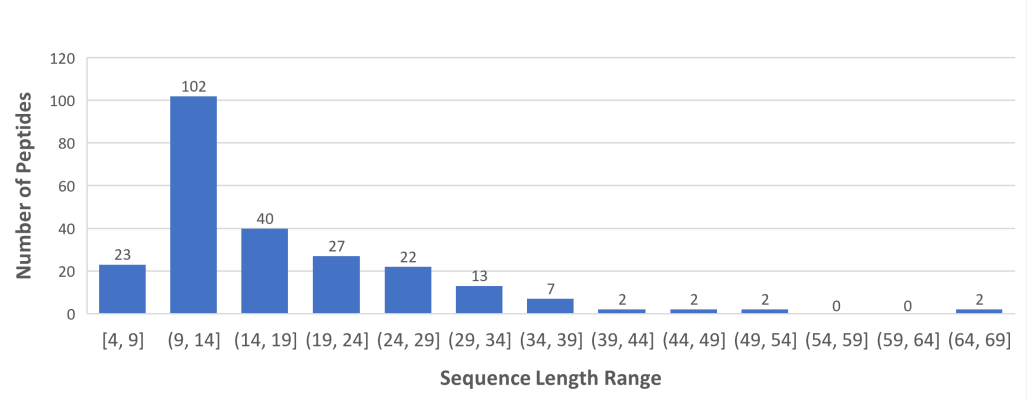

Fig. 6. Distribution of antibiofilm peptides in various sequence length.

4.1.2 MBIC Value Distribution

The number of peptides in different MBIC value ranges are plotted in Fig. 7. The graph shows that almost 69% of peptides have an MBIC value lower than 11 μM. The lower the MBIC value, the more influential the peptide will be against biofilm. Our curated data showed that only a few peptides have values of more than 128 μM, and those peptides are less effective against biofilm. Hence we avoided those peptides in our regression model.

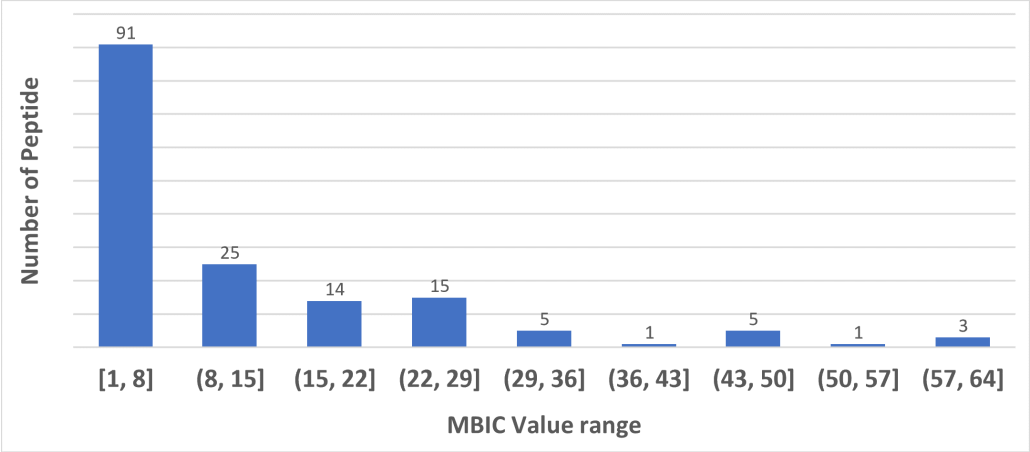

Fig. 7. Distribution of MBIC value of antibiofilm peptides in the positive dataset.

4.1.3 Amino Acid Composition Analysis

To compare the presence of various amino acids in the positive and negative set, we used the average percentage composition of the amino acid against antibiofilm peptides and non-antibiofilm peptides (curated for model B). Fig. 8 shows the comparison. An asterisk (*) in the figure denotes statistical significance ($P < 0.05$). Antibiofilm peptides have high percentages of lysine (K) and arginine(R). Lysine and Arginine are positively charged amino acids. Also, the ratio of leucine (L), isoleucine (I), phenylalanine (F), and tryptophan (W) are high in antibiofilm peptides. Leucine is mainly responsible for the α-helix formation, while phenylalanine and tryptophan are aromatic amino acids. The non-antibiofilm peptides have more percentage of aspartic acid (D), glutamic acid (E) and methionine (M). While aspartic acid and glutamic acid are highly negatively charged amino acids, methionine is of neutral charge. The comparison between two kinds of peptides shows antibiofilm peptides are high in positively charged and aromatic amino acids. That is the reason antibiofilm peptides are effective against negatively charged bacterial membranes. We did two tailed t-test using microsoft excel for analysis of the

data. Our statistical analysis showed a significant difference ($P<0.05$) between the average percentage of arginine and lysine in antibiofilm and non-antibiofilm peptides.

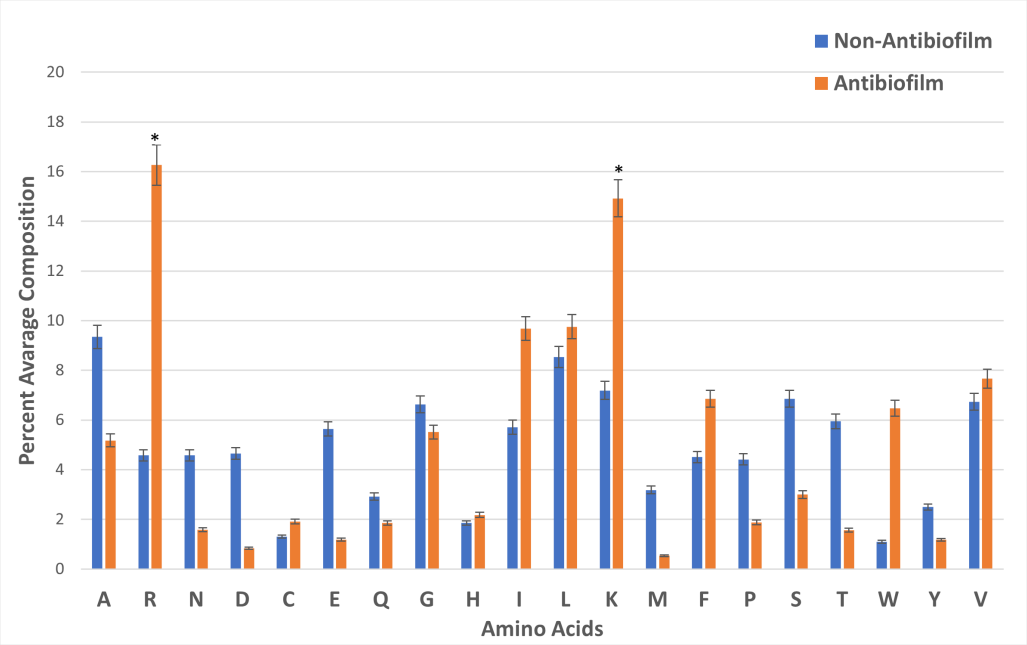

Fig. 8. Distribution of average amino acids percentage composition in antibiofilm and non-antibiofilm peptides. Error bars show standard error of the mean.

4.1.4 Secondary Structure Analysis

The secondary structure of the peptides was also being compared for antibiofilm and non-antibiofilm peptides. The average percentage composition of secondary structure showed a higher percentage of α-helix in antibiofilm peptides, while non-antibiofilm peptides are higher in coil formation. The higher helical property allows the antibiofilm peptide to fold into the amphipathic structure, which helps them to disrupt membrane activity. A two tailed t-test analysis showed a significant difference between the mean value of antibiofilm and non-antibiofilm peptides when comparing different secondary structures. Fig. 9 shows the comparison. An asterisk (*) in the figure denotes statistical significance ($P < 0.05$).

4.1.5 Hydrophobicity Analysis

The comparison of hydrophobic properties between antibiofilm peptides and non-antibiofilm peptides showed the higher amount of hydrophobicity of the antibiofilm peptides. The hydrophobic portion of antibiofilm peptides leads to insertion of the peptides into the less polar bacterial membrane and leads to destabilizing membrane barriers [43]. Fig. 10 shows the comparison. An asterisk (*) in the figure denotes statistical significance ($P < 0.05$). The higher percentage of alanine, valine, leucine, isoleucine, and phenylalanine could be a potential reason for the antibiofilm peptides' hydrophobic nature. A two tailed t-test analysis of hydrophobic, hydrophilic, and neutral properties showed a significant difference between antibiofilm and non-antibiofilm peptides.

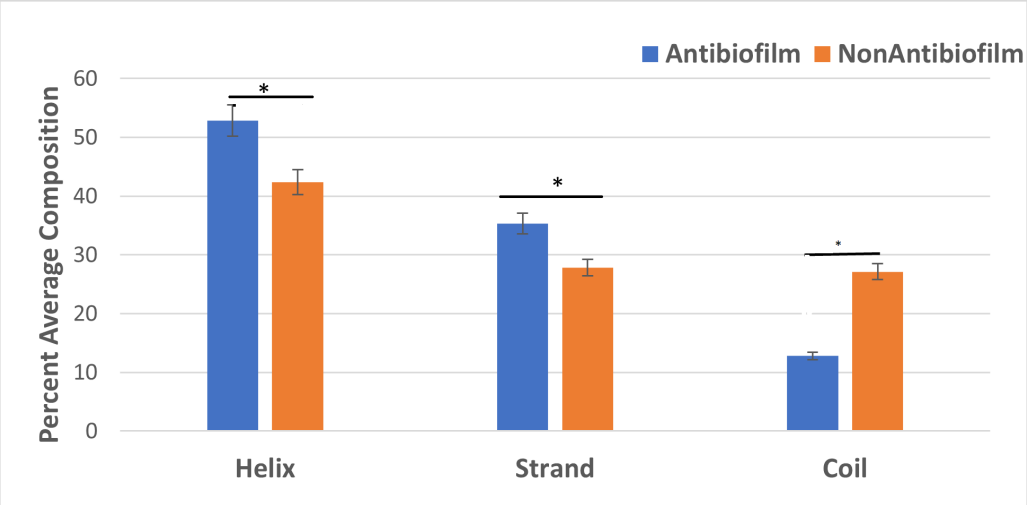

Fig. 9. Distribution of average secondary structure in antibiofilm and non-antibiofilm peptides. Error bars show standard error of the mean.

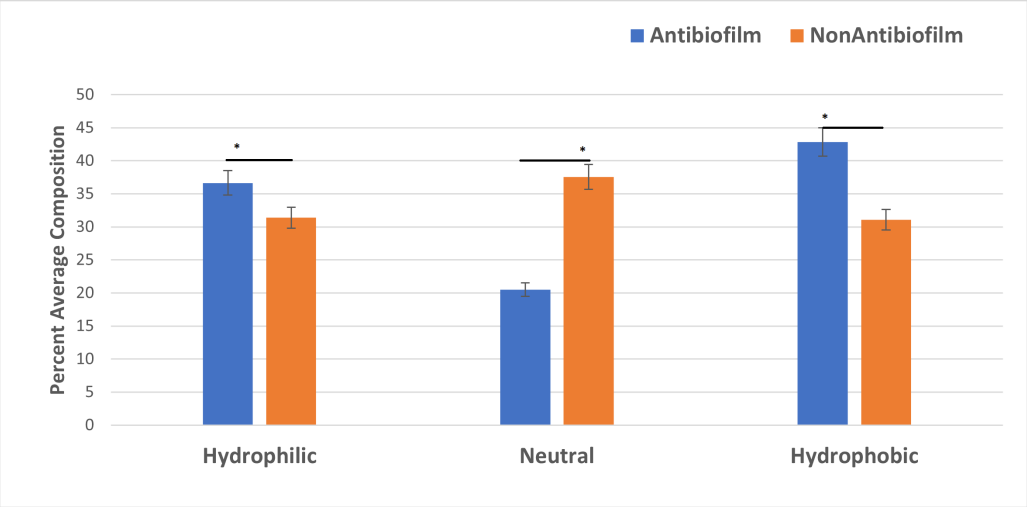

Fig. 10. Distribution of different hydrophobic properties in antibiofilm and non-antibiofilm peptides. Error bars show standard error of the mean.

4.1.6 Motif Analysis

We also compared the motif present in antibiofilm and non-antibiofilm peptides. The presence of the motif is listed in Table 2. The motif analysis showed the highest number of motif present in the antibiofilm peptides is: 'R I R V.' Also, motifs like 'R I V Q R I K, "R I V Q R,' 'K R I V Q R I K,' 'I G K E F K R' have a higher presence in antibiofilm peptides. In non-antibiofilm peptides, the most abundant motifs are 'S D,' 'L E,' 'S E,' and 'E P.'

4.1.7 Different MBIC Value Analysis

To see any specific correlation between lower MBIC values and the physio-chemical properties of peptides, we compared two groups of peptides from our MBIC dataset. We considered around 44 peptides, which have MBIC value lower or equal to 4 μM in group 1 and another 10 peptides with a high MBIC value greater than 124 μM in group 2. We assessed the average percentage of different properties like the percentage of alanine, the percentage of lysine, etc. Fig. 11 shows the difference between the two groups of peptides.

Table 2
Top 5 Motif Patterns and Number of Occurrences in Positive and Negative Dataset

Peptide Type	**Motifs**	**Number of Occurrences**
Antibiofilm	R I R V	19
Antibiofilm	R I V Q R	14
Antibiofilm	R I V Q R I K	13
Antibiofilm	I V Q R I K	13
Antibiofilm	G K E F K R	12
Non-Antibiofilm	S D	95
Non-Antibiofilm	L E	89
Non-Antibiofilm	S E	87
Non-Antibiofilm	E P	86
Non-Antibiofilm	E D	81

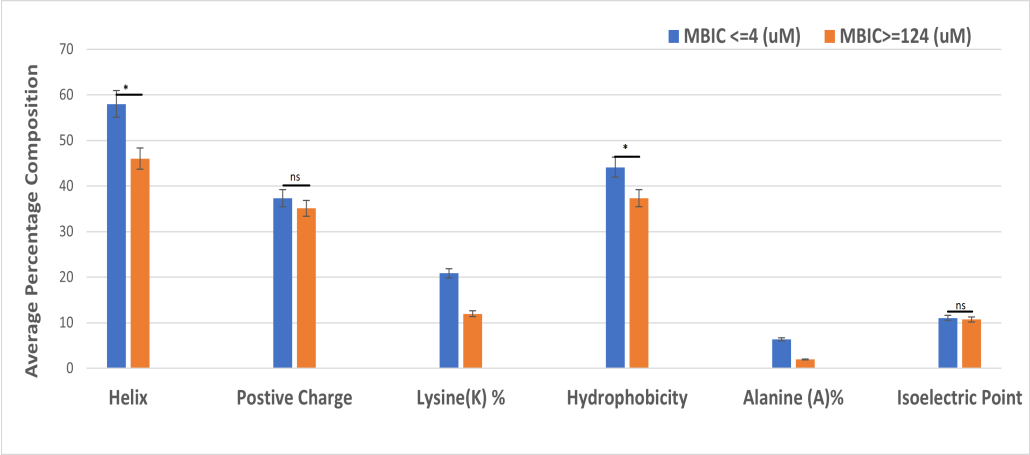

Fig. 11. Comparison between different properties of peptide grouped by MBIC value. Error bars show standard error of the mean.

While we noticed an increasing spike for the helical structure in group 1 peptides, we could not find much difference in the isoelectric point between them. Also, a higher percentage of alanine (A) and lysine (K) in group 1 peptides could contribute to the helical property and the effectiveness of peptides. We noticed from our two tailed t-test that helix and hydrophobicity showed statistical significance with $P < 0.05$. The isoelectric point and positive net charge did not show any significant statistical difference.

4.2 Performance Analysis Of Different Machine Learning Model

With the popularity of machine learning, several works have been done on peptide prediction using different machine learning algorithms. Using machine learning to detect a novel peptide not only saves time and cost but also gives an immense opportunity to combine various peptide features to predict an accurate result. In our work, we used different features to develop our machine learning models. We also used multiple machine learning algorithms to evaluate the best model against our dataset and features. We worked with a combination of different features like AAC and Motif, DPC and Motif, CTD and Motif. The details can be found in appendix section (A). According to Gupta et al., combining DPC with Motif features and running them with the SVM, gave the best result. Hence we refer to that approach as our 'baseline' method for the rest of the thesis. In our work, we found that the best performance could be achieved while combining AAC, DPC, CTD, and motif features together. We refer this approach as our 'implemented' approach for the rest of the the thesis. The detailed performance analysis of the model is explained in our next section.

4.2.1 Performance For Model A (Working With The Main Dataset)

As our classification model is a binary classifier, we emphasized the F1 and MCC values to judge our models' accuracy. For our Model 'A,' we used only randomly generated data, as mentioned beforehand. The model showed the best performance with the 'implemented' approach by applying the SVM algorithm compared to RF, XGBoost,

or MLPClassifier. We initially tried to work with a 'linear' kernel-mode, which did not show a satisfactory performance due to the nature of our dataset. We also tried recursive feature elimination with linear kernel SVM but using 'RBF' kernel outperformed other methods. The SVM model with the 'RBF' kernel achieved an F1 score of 0.9247 while the MCC was 0.9181 on the validation dataset. This model gave an accuracy score of 98.68%, sensitivity 89.58%, and specificity 99.58% on the same validation dataset. Our baseline approach could only achieve an F1 score of 0.8484, MCC of 0.8334, accuracy 97.24%, sensitivity 87.5%, and specificity 98.14% using. We reached the average cross-validation accuracy of 0.9793 (with Standard deviation 0.0047), F1 score 0.8741 (with Standard deviation 0.0330), MCC score 0.8689 (with Standard deviation 0.0324) in the training dataset using SVM. The performance of different models on the validation dataset is given in the Table 3. The specificity, sensitivity and accuracy in table are listed in percentage (%) format.

4.2.2 Performance For Model B (Working With Alternative Dataset)

We evaluated performance between different machine learning algorithms for the alternative dataset as we did for Model A. We assessed the 'baseline' approach as well as our 'implemented' approach on this dataset. We applied SVM, random forest, XGBoost, and MLPCassifier also on the alternative dataset. The SVM model with the 'RBF' kernel gave the best performance. The model B achieved an F1 score of 0.9090, an MCC value of 0.9054, an accuracy of 98.68%, a sensitivity of 83.33%, and a specificity of 100% on the validation dataset. We reached the average cross-validation accuracy of 0.9727 (with Standard deviation 0.0058), F1 score 0.8319 (with Standard deviation 0.0393), MCC score 0.8249 (with Standard deviation 0.0410) on the training dataset using SVM. The performance of different models on the validation dataset is given in the Table 4. The specificity, sensitivity and accuracy in table are listed in percentage (%) format.

Table 3
Performance Evaluation of Different Machine Learning Techniques with Model A

Model Performance	**Specificity**	**Sensitivity**	**Accuracy**	**F1 Score**	**MCC**
Baseline Approach	98.14	87.5	97.24	0.8484	0.8334
SVM with rbf kernel, gamma = 0.1	**99.58**	**89.58**	**98.68**	**0.9247**	**0.9181**
Random Forest (n-estimators = 100)	100	70.83	97.37	0.8292	0.8297
XGBoost (n-estimators = 50, max-depth = 10)	99.17	85.41	97.93	0.8817	0.8709
MLPClassifier (hidden-layer-sizes = (100, 100, 100)	98.55	89.58	97.74	0.8775	0.8653

4.2.3 Comparison With Existing Models

There are only very few antibiofilm peptide prediction models currently available compared to other antimicrobial peptides prediction models. We used the Gupta et al. dataset to compare our model because it has more data points than other existing models. Our implemented approach involved a combination of different features like AAC, DPC, CTD, and motifs. Our extensive feature set could work correctly on more massive datasets, as we found in Gupta et al. We extracted the positive and the negative peptides, ran the training data on our model and compared the validation dataset result with the best result reported by Gupta et al. Our model could achieve an MCC score of 0.9050

Table 4
Performance Evaluation of Different Machine Learning Techniques with Model B.

Model Performance	**Specificity**	**Sensitivity**	**Accuracy**	**F1 Score**	**MCC**
Baseline Approach	98.35	85.41	97.18	0.8453	0.8229
SVM with rbf kernel, gamma =0.1	**100**	**83.33**	**98.49**	**0.9090**	**0.9054**
Random Forest (n-estimators=100)	99.79	66.66	96.81	0.7901	0.7894
XGBoost (n-estimators=50, max-depth=10)	97.56	75.41	97.56	0.8631	0.8498
MLPClassifier (hidden-layer-sizes=(100, 100, 100)	98.55	83.33	97.18	0.8421	0.8267

compared to the reported MCC of 0.84. Our model could also achieve an accuracy of 98.46%, a sensitivity of 86.11%, and a specificity of 99.71%. The F1 score of our model was 91.17, while we could not find an F1 score mentioned in the paper. The comparison result can be found in Table 5. The specificity, sensitivity and accuracy in table are listed in percentage (%) format.

4.2.4 Performance Of Regression Model

The effectiveness of antibiofilm peptides can be evaluated by the minimum biofilm inhibitory concentration (MBIC). The antibiofilm peptides with lower MBIC value are considered more effective and assessed in this scope of work. We started building our model with the linear regression algorithm, but this algorithm could not perform well due

Table 5
Performance Comparison of Our Method with the Existing Record

Validation dataset performance	**Specificity**	**Sensitivity**	**Accuracy**	**F1 Score**	**MCC**
Reported in paper	97.75	91.67	97.19	Not Known	0.84
Achieved with our model	**99.71**	86.11	**98.46**	**91.17**	**0.90**

to the nature of the dataset. Instead, we started working with different non-linear regression algorithms. We found that the support vector regression (SVR) and XGBRegression (XGBR) worked better to predict the MBIC value. But the linear kernel SVR with recursive feature elimination could not perform well. Then, we used the 'RBF' kernel for the SVR, which outperformed others. After many trial and error and cross-validation, we fixed the 'C' value as 10 and gamma as 0.1. We divided the dataset into a 90:10 ratio and performed ten-fold cross-validation. The performance was measured based on the root mean squared error (RMSE) and the standard deviation. These error metrics are directly related to the correlation coefficient of the model. SVR model achieved the best RMSE value of 4.7786 and a standard deviation of 1.5085 on the test dataset. Table 6 shows the performance of different models.

The predicted MBIC values from the SVR with the 'RBF' kernel model are plotted against the original values of the test dataset in Fig. 12.

Table 6
Performance Comparison of Different Regression Models

Model	**RMSE**	**Standard Deviation**
SVR, rbf kernel, C=10, gamma=0.1	**4.7786**	**1.5085**
XGBR, n-estimators=5, max-depth=7	6.2285	3.7339
SVR, linear kernel, 100 feature	8.9225	7.0579

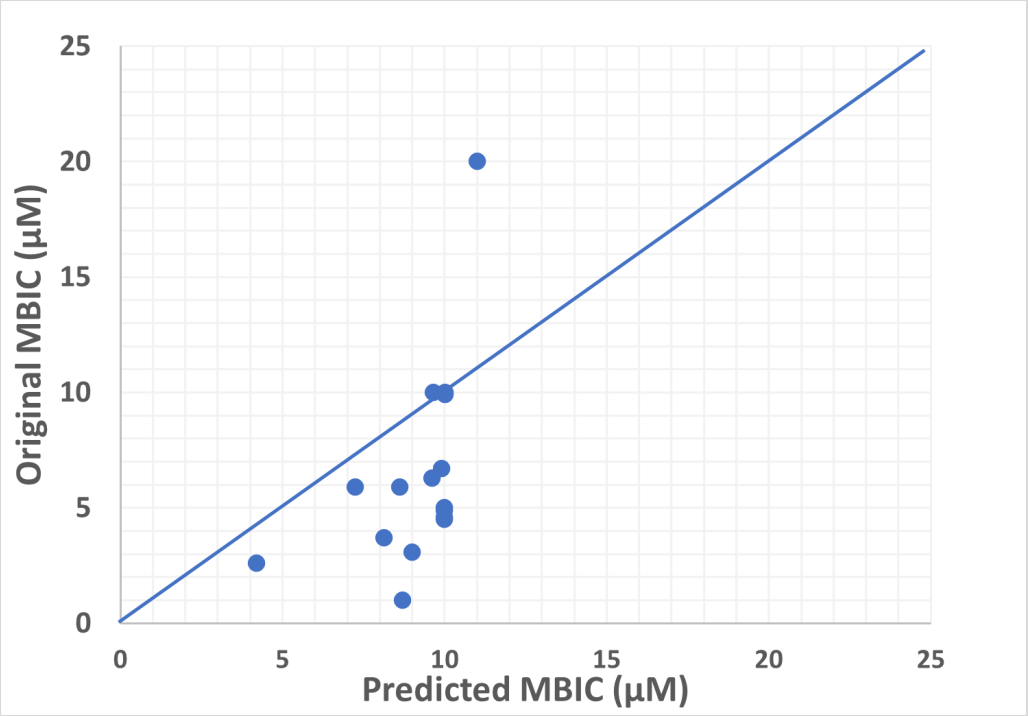

Fig. 12. The distribution of original and predicted MBIC value from SVR Model.

4.3 Prediction Of Antibiofilm Peptides

To see the presence of antibiofilm properties in peptides from diverse habitats, we collected different sets of peptides. We ran our classification model, regression analysis,

and motif analysis on probable peptides to see the possibilities of having antibiofilm properties. The steps of having a potential antibiofilm candidate is given in Fig. 2. We selected a few antibiofilm peptides with high prediction probability from the classifier and then ran the regression analysis to determine the effectiveness. We also checked the motif counts of the peptides to see the similarity with our antibiofilm dataset. We ranked the peptides as an effective antibiofilm candidate depending on the higher percentage of classification probability, a higher value of motif, and lower MBIC prediction value (shown in Fig. 2). As we achieved a better performance result with our model A (SVM with the main dataset), we used that model to classify the peptides while the SVR model predicted the MBIC value.

4.3.1 Prediction With Anticancer Peptides

We assessed a test dataset of 74 anticancer peptides from the DRAMP database. We evaluated this dataset in our classification model. Seventeen peptides among them showed higher prediction (>96%) percentages from the classifier model. Then we ran the motif analysis for those 17 peptides and regression analysis for the MBIC value. After our research, we identified top 10 peptides that were evaluated further for potential antibiofilm effects. Table 7 shows the details of the peptides that we considered for validation in the lab.

4.3.2 Prediction With Antiviral Peptides

We secured a test dataset of 213 antiviral peptides from the DRAMP database. We examined this dataset in our classification model. Eleven peptides showed a higher probability prediction (>96%) percentage from the classifier model. Then we ran the motif analysis and regression analysis for those eleven peptides. After our study, we developed a top-10 peptide list that could be evaluated further for potential antibiofilm effects. Table 8 shows the details of the peptides that we considered for further evaluation.

Table 7
The Evaluation of Anticancer Peptides for Potential Antibiofilm Activity

Name	**Prediction Probability from Classifier**	**Predicted MBIC from Regression (μM)**	**Motif Count***
DRAMP03574	0.98384803	15.21	89
DRAMP02926	0.99259543	13.6	63
DRAMP02990	**0.99993214**	**5.67**	**54**
DRAMP03575	0.99662332	14.3	54
DRAMP02989	**0.99985197**	**5.65**	**13**
DRAMP03829	**0.999935172**	**5.01**	**10**
DRAMP18494	**0.99522187**	**7.97**	**10**
DRAMP03687	0.99412033	12.06	8
DRAMP01110	0.9650794	12.07	7
DRAMP04133	0.993333	8.69	6
*The peptides are sorted based on motif count.			

Table 8
The Evaluation of Antiviral Peptides for Potential Antibiofilm Activity

Name	**Prediction Probability from Classifier**	**Predicted MBIC from Regression (μM)**	**Motif Count***
DRAMP02926	**0.99259543**	**13.61810654**	**63**
DRAMP01110	0.965079491	12.07187072	7
DRAMP02761	**0.999989229**	**8.93296784**	**6**
DRAMP18688	0.98513251	11.75484474	4
DRAMP04506	0.975836702	8.192149472	4
DRAMP01084	0.994488209	11.75073392	2
DRAMP01086	0.993796461	13.06669082	2
DRAMP04010	0.981415481	13.11930625	2
DRAMP02992	0.991855769	8.035789535	1
DRAMP02233	0.973355257	13.13148776	1
*The peptides are sorted based on motif count.			

4.3.3 Prediction with MilkAMP Database

We evaluated a test dataset of around 300 peptides from the MilkAMP database. Sixty six peptides showed higher prediction (>96%) percentage from the classifier model than the previous two cases. We ran the motif analysis and regression analysis for those 66 peptides. Regression analysis showed a comparatively higher MBIC value than the previous two cases. After our study, we identified a top-10 peptide list (Table 9), which could be evaluated further for potential antibiofilm characteristics.

Table 9
The Evaluation of Milk Peptides for Potential Antibiofilm Activity

Name	**Prediction Probability from Classifier**	**Predicted MBIC from Regression (μM)**	**Motif Count***
LFB0093	**0.990627**	**12.82877**	**246**
LFB0091	**0.988824**	**12.91982**	**246**
LFB0155	0.992153	13.79659	169
LFB0177	0.981204	12.55854	167
LFB0176	0.978091	13.2394	167
LFB0173	0.988314	11.68185	108
LFB0134	0.995092	13.39625	104
LFB0135	0.995497	13.35955	103
LFB0131	0.994338	13.40014	103
LFB0146	0.993795	19.75503	103
*The peptides are sorted based on motif count.			

4.3.4 Prediction With Antimicrobial Peptides

We also worked with a large dataset of antimicrobial peptides from the DRAMP database. As antibiofilm is a subset of antimicrobial peptides, we removed the similar peptides from the dataset. After removing duplicates, around 4700 peptides remained to be evaluated as a test set. When we ran this large set of data with our model, as expected, we received a higher number of peptides compared to the other cases from the classifier.

We assessed this large antimicrobial dataset to see similarities with antibiofilm functionality in diverse habitats like antibacterial or antifungal peptides. After running the regression model and motif analysis, we listed the top 10 peptides for a potential antibiofilm candidate. The details of the evaluated peptides can be found in Table 10.

Table 10
The Evaluation of Antimicrobial Peptides for Potential Antibiofilm Activity

Name	**Prediction Probability from Classifier**	**Predicted MBIC from Regression (μM)**	**Motif Count***
DRAMP04014	0.987356462	15.33777	336
DRAMP04015	0.985853713	15.59084	335
DRAMP02648	0.996209262	16.67575	319
DRAMP03645	0.956063771	13.06714	294
DRAMP03570	0.999995939	11.07888	205
DRAMP02709	0.969905986	14.82536	175
DRAMP18616	**0.999999824**	**7.999017**	**171**
DRAMP18617	**0.999993843**	**6.723098**	**140**
DRAMP02520	**0.989761244**	**5.842829**	**138**
DRAMP02365	**0.97699594**	**8.344172**	**135**
*The peptides are sorted based on motif count.			

5 DISCUSSION

With the popularity of machine learning, the literature shows an increasing number of computational models and databases related to peptides. However, there is less work on antibiofilm peptides modeling compared to antimicrobial peptides. The main reason could be the lower availability of antibiofilm peptides data due to their complicated characteristics. Our goal was to develop a classification model to find antibiofilm features and generate a prediction model to understand the efficacy of the peptides. We did not just restrict the performance of our model against the randomly generated negative peptide dataset. We also evaluated the performance of our model against peptides that already exist in nature. We concluded that the model performance could change when we change the negative dataset. The performance of our model against the main and additional database also showed variation in accuracy percentage. Our model performed better with randomly generated peptides. However, the current negative dataset could be replaced with peptides, which are proven to form a biofilm and evaluated against our model.

We also considered stratified sampling and ten-fold cross-validation to eliminate the overfitting problem due to an imbalanced dataset. The training and testing performance metrics showed a clear indication that our model did not overfit. While our model achieved the best performance with SVM, the other methods, such as XGBoost or Recursive Feature Analysis, might also do well with more hyperparameter tuning. We do not currently have a vast dataset, but MLPClassifier could work better than SVM due to its training capacity with an increased dataset.

Our model achieved better performance with the same data from Gupta et al. but did not perform well with the dataset of Sharma et al. [13] and Fallah Atanaki et al [15]. Our model could reach only the F1 score of 0.8965 against the dataset used by Fallah Atanaki et al. The reason could be the dataset variation between our model and theirs. Our model is built on the idea that the antibiofilm peptides are rarer in nature than other peptides. To

mimic that concept, we used ten folds more peptides in negative data than the positive dataset. Fallah Atanaki et al. developed a model in which the positive dataset size was more than two times larger than the negative dataset. The basic assumption of building the machine learning model is different in these two cases, which could lower the performance of our model for that dataset.

While developing our feature set, we used a different approach in motif analysis than seen in the literature. Most of the time, the total positive dataset is evaluated for unique motifs against the negative dataset, and the result is fed to the model. Per the MERCI software, we can find motifs that are present in the positive dataset and absent in the negative dataset. Providing this 'privileged' information to the model may predict a higher accuracy of the model. To eliminate this situation, we searched motifs for each training set while running our cross-validation. We stored the motifs of the training dataset. Then we examined the motif on the test dataset from the stored motif of the training set. Also, we found all the default motifs by tuning specific parameters to enhance the effect of motif features on our dataset. Increasing the limit of motif count changed the performance compared to the 'without motif' model.

Our data analysis showed a higher percentage of positively charged amino acids like lysine and arginine in antibiofilm peptides. The higher positive net charge and higher hydrophobicity are characteristic of antibiofilm activity peptides [10]. Though we saw a higher percentage of α-helix, the overall peptide structure showed a varied percentage of β-sheets and coils. Our data analysis showed a significant difference between antibiofilm and non-antibiofilm peptides when analyzing the properties mentioned above.

Though our work does not have much scope to analyze each antibiofilm peptide and its design, we studied human cathelicidin LL-37 and some of its derived peptides. The difference in sequence could change the effectiveness of these peptides against biofilm formed by *Pseudomonas aeruginosa*. The LL-37 prevents biofilm formation of

P.aeruginosa in a concentration lower than its MIC value by probably blocking the growth of extracellular matrix [44]. The LL19-37 could not affect biofilm, but adding a group of 'I G K E F K' (LL13-37) changed peptide effectiveness. LL13-37 shows the inhibition of biofilm formation at 5μM. 'I G K E F K' is one of the motifs we found in high numbers in our positive dataset during motif analysis. While LL-19 has no activity against bacterial membrane permeability [44], adding a motif of 'I V Q R I K' increases permeability significantly in LL-25. 'I V Q R I K' is another motif that we found in our positive dataset.

While building the regression model, we discovered that all the peptides may not have significant MBIC value. We started our data collection to see the eradication efficacy against a preformed biofilm (minimum biofilm eradication concentration, MBEC). Due to the data availability, we had to use only the MBIC value as a parameter for efficacy. MBEC and MBIC values are mostly different. Furthermore, antibiofilm peptide, which has a lower MBIC value, may not effectively eradicate preformed biofilm. Dermaseptin-AC is very useful in the inhibition of biofilm (MBIC value 32 μM) formed by *Staphylococcus aureus*. This peptide is not effective in eradicating preformed biofilm (MBEC value 256 μM) [45]. We did not find any computational model so far to predict the MBIC value for an antibiofilm peptide. Thus our regression model is a new approach to assess the efficacy. In our regression model, we considered a smaller dataset with an MBIC value <= 64 μM. We did not consider the higher MBIC value to eliminate the outliers. Our regression model could be biased and might not be effective against an unknown peptide with a higher MBIC value. Also, the MBIC values were curated against different pathogens due to the unavailability of data. Collecting data against the same pathogen could increase the reliability of the model. The recursive feature elimination on the regression model could improve the performance of the model. We considered our regression model as a ranking parameter to judge the potential antibiofilm peptide. This

prediction will help to rank thousands of predicted peptides by classifier when run against a huge peptide database. We gave motif analysis a significant weightage, as we already saw in the literature [13].

To find potential biofilm peptides from diverse habitats, we considered different peptide datasets to evaluate antibiofilm activity. We analyzed peptides like antimicrobial, antiviral, anticancer, etc. Literature shows that antimicrobial peptides may be viewed as a negative dataset while considering anticancer peptides [46]. But we also saw examples of elements like bioactive selenium compounds that could be anticancer and effective on biofilm [47]. That is why we evaluated anticancer peptides for biofilm activity. We further studied the top 10 anticancer peptides, predicted by our model for antibiofilm activity. Further research showed that many peptides contain higher percentages of positively charged lysine (K) and arginine (R), for example, DRAMP02990, DRAMP02989, DRAMP18494, and DRAMP01110. Most of the peptides are potentially active against planktonic bacteria. We also evaluated the minimum inhibitory concentration (MIC) value of those peptides. Our research showed that DRAMP02990 has a MIC value of 1.4 μM against *E. Coli*, while DRAMP01110 has a MIC of 2.7 μg/ml against *E.Coli*. DRAMP18494 has a MIC value of 8 μg/ml against *Staphylococcus aureus*, and DRAMP02989 has a MIC value of 14.3 μM. The regression model also predicted a lower MBIC ($<=$ 10 μM) for all these four peptides. So we ranked these four peptides higher even in the top ten candidates for potential *in-vitro* analysis.

There are chances that viruses could infect bacteria and kill them or make them less effective for biofilm formation. Current literature also shows that engineered peptides could be effective against both biofilm and viruses [48]. We evaluated the efficacy of antiviral peptide for antibiofilm activity. We further assessed the top 10 candidates from the antiviral peptide list. Most of them have a lower motif count. DRAMP02926 has the highest motif count and also is a potential anticancer peptide. We could not find any MIC

value for this peptide and the predicted MBIC value is also high. However, the other peptide, DRAMP02761, has a better chance for potential antibiofilm properties. DRAMP02761 is effective against gram-positive and gram-negative microbes.

There is also a possibility that raw milk can preserve the antibiofilm property of peptide [49]. To evaluate any potential dairy peptide that can be effective against biofilm, we worked with the milk database. Further analysis of the top 10 peptides from Table 9 revealed that both the peptides LFB0093 and LFB0091 have a MIC value of 6μM against *E.Coli*.

When we further analyzed the peptides from the antimicrobial dataset, we found some interesting facts. We already know human cathelicidin LL-37 is an effective antibiofilm peptide. Our model found out Gorilla cathelicidin (DRAMP02709) as a potential candidate for biofilm activity. This peptide belongs to the same 'Cathelicidin Antimicrobial Peptide' (CAMP) gene family and has a high percentage of cationic charge amino acids, lysine and arginine. Our model also pointed to peptides like DRAMP18617 and DRAMP18616 with adequate killing capacity against gram-negative bacteria (MIC 4.0 μM and 6.0 μM). Both the peptides are synthetic derivatives of a well-known antibiofilm peptide SMAP-29. Our model also indicated DRAMP02365 as antibiofilm peptide. DRAMP02365 is similar to Pleurocidin and active against both gram-positive and gram-negative bacteria (*E. coli* (MIC=1 μg/ml, *S. aureus* (MRSA) C623 (MIC=16 μg/ml)).

Having a significant efficacy against planktonic bacteria (lower MIC value) does not guarantee antibiofilm activity. However, the antimicrobial activity of a new peptide certainly increases the chance of effectiveness against biofilm. We believe our approach to this computational model is different from the literature, and the model can work well in the real world.

6 FUTURE WORK

There are a few essential things to consider for our future work. Due to time constraints and the unavailability of resources, our current work scope did not permit us to obtain them. The *in vitro* evaluation of the predicted peptides from our model using a bacterial assay in a laboratory setting would be the best validation part of our model.

If we can collect effectiveness of peptides, i.e., the MBIC data against a specific pathogen with larger samples, we could build a robust regression model with a lower RMSE score. We would also like to build a classifier in our pipeline to predict if the peptide falls under the group of lower MBIC value (<= 64 μM). The peptides with higher value would not be considered for further analysis. The peptides, which will be classified as having a probability of lower MBIC value, could be further analyzed with the regression model. This process would help to eliminate any biased predictions.

Evaluating our model with natural peptides rather than randomly generated peptides will be exciting work to do. We want to assess the probability of antibiofilm activity in different species and genomes. We are looking forward to working with some metagenomic datasets from different habitats like sea-water or soil microbes. With a larger dataset, we could apply other machine learning techniques like neural networks to improve model performance and efficacy.

7 CONCLUSION

Biofilm is one of the main reasons for causing chronic and implant-related infection. The growing resistance of biofilm against well known antibiotics makes it an essential topic for research. While biofilm's mechanism of action is yet not fully discovered, there is a growing need to find a solution against these colonies of pathogens. Our research work focused on the classification of antibiofilm peptides and predicting the efficacy of those peptides from diverse habitats. The model was built after evaluating many important peptide features and curating data from the literature. The research work suggests a cost-effective approach to developing machine learning models to deal with the situation when a limited dataset and resources are available. The work also provides a vast scope to evaluate a much larger dataset than an *in vitro* approach. The model performance looks very promising. The unique way of ranking the top candidate for the antibiofilm activity will lead to a faster validation process in the laboratory.

Literature Cited

[1] C. for Disease Control and Prevention., "Antibiotic-resistant germs: New threats. (2020, october 28).."

[2] J. Tanwar, S. Das, Z. Fatima, and S. Hameed, "Multidrug Resistance: An Emerging Crisis," *Hindawi*, vol. 2014, no. 541340, 2014.

[3] L. S. of Hygiene & Tropical Medicine, "United nations and who call for urgent action to avert antimicrobial resistance crisis - expert comment."

[4] Q. Wu, J. Patočka, and K. Kuča, " Insect Antimicrobial Peptides, a Mini Review.," *Toxins*, vol. 10, no. 461, 2018.

[5] M. Kostakioti, M. Hadjifrangiskou, and S. Hultgren, " Bacterial Biofilms: Development, Dispersal, and Therapeutic Strategies in the Dawn of the Postantibiotic Era.," *Cold Spring Harbor Perspectives in Medicine*, vol. 3(4), no. a010306, 2013.

[6] N. Høiby, O. Ciofu, and B. T., " Pseudomonas aeruginosa biofilms in cystic fibrosis. ," *Future Microbiol*, vol. 5(11), no. 1663-74, 2010.

[7] Z. Khatoon, C. D. McTiernan, E. J. Suuronen, T.-F. Mah, and E. I. Alarcon, " Bacterial biofilm formation on implantable devices and approaches to its treatment and prevention. ," *Heliyon*, vol. 4(12), no. e01067, 2018.

[8] V. R., " Biofilms: microbial cities of scientific significance.," *J Microbiol Exp.*, vol. 1(3), pp. 84–98, 2014.

[9] A. Di Somma, A. Moretta, C. Canè, A. Cirillo, and A. Duilio, " Antimicrobial and Antibiofilm Peptides," *Biomolecules*, vol. 10(4), no. 652, 2020.

[10] D. Pletzer and R. E. W. Hancock, "Antibiofilm peptides: Potential as broad-spectrum agents," *Journal of Bacteriology*, vol. 198, no. 19, pp. 2572–2578, 2016.

[11] M. Di Luca, G. Maccari, G. Maisetta, and G. Batoni, " BaAMPs: the database of biofilm-active antimicrobial peptides.," *Biofouling*, vol. 31(2), p. 193–199, 2015.

[12] G. Wang, X. Li, and Z. Wang, "APD3: the antimicrobial peptide database as a tool for research and education," *Nucleic Acids Research*, vol. 44, pp. D1087–D1093, 11 2015.

[13] S. Gupta, A. K. Sharma, S. K. Jaiswal, and V. K. Sharma, "Prediction of biofilm inhibiting peptides: An in silico approach," *Frontiers in Microbiology*, vol. 7, p. 949, 2016.

[14] A. Sharma, P. Gupta, R. Kumar, and A. Bhardwaj, " dPABBs: A Novel in silico Approach for Predicting and Designing Anti-biofilm Peptides.," *Scientific Reports*, vol. 6, no. 21839, 2016.

[15] F. Fallah Atanaki, S. Behrouzi, S. Ariaeenejad, A. Boroomand, and K. Kavousi, "Bipep: Sequence-based prediction of biofilm inhibitory peptides using a combination of nmr and physicochemical descriptors," *ACS Omega*, vol. 5, no. 13, pp. 7290–7297, 2020.

[16] V. Machado, J. Gelinski1, C. M. Baratto1, E. M. Borges2, V. A. Vicente3, M. M. F. Nascimento, and G. G. Fonseca, " Technological Potential of Antimicrobial Peptides: a Systematic Review.," *The Indian Journal of Pharmaceutical Sciences*, vol. 81, pp. 807–814, 2019.

[17] M. Dostert and R. E. W. Belanger, C. R.and Hancock, " Design and Assessment of Anti-Biofilm Peptides: Steps Toward Clinical Application.," *Journal of Innate Immunity*, vol. 11, p. 193–204, 2019.

[18] M. R. Yeaman and N. Y. Yount, "Mechanisms of antimicrobial peptide action and resistance," *Pharmacological Reviews*, vol. 55, no. 1, pp. 27–55, 2003.

[19] M. Pasupuleti, A. Schmidtchen, and M. Malmsten, "Antimicrobial peptides: key components of the innate immune system," *Critical Reviews in Biotechnology*, vol. 32, no. 2, pp. 143–171, 2012.

[20] E. Sun, C. R. Belanger, E. F. Haney, and R. E. Hancock, "10 - host defense (antimicrobial) peptides," in *Peptide Applications in Biomedicine, Biotechnology and Bioengineering* (S. Koutsopoulos, ed.), pp. 253 – 285, Woodhead Publishing, 2018.

[21] R. Kapoor, M. W. Wadman, M. T. Dohm, A. M. Czyzewski, A. M. Spormann, and A. E. Barron, "Antimicrobial peptoids are effective against pseudomonas aeruginosa biofilms," *Antimicrobial Agents and Chemotherapy*, vol. 55, no. 6, pp. 3054–3057, 2011.

[22] K.-i. Okuda, T. Zendo, S. Sugimoto, T. Iwase, A. Tajima, S. Yamada, K. Sonomoto, and Y. Mizunoe, "Effects of bacteriocins on methicillin-resistant staphylococcus

aureus biofilm," *Antimicrobial Agents and Chemotherapy*, vol. 57, no. 11, pp. 5572–5579, 2013.

[23] J. Overhage, A. Campisano, M. Bains, E. C. W. Torfs, B. H. A. Rehm, and R. E. W. Hancock, "Human host defense peptide ll-37 prevents bacterial biofilm formation," *Infection and Immunity*, vol. 76, no. 9, pp. 4176–4182, 2008.

[24] C. de la Fuente-Núñez, F. Reffuveille, E. F. Haney, S. K. Straus, and R. E. W. Hancock, "Broad-spectrum anti-biofilm peptide that targets a cellular stress response," *PLOS Pathogens*, vol. 10, pp. 1–12, 05 2014.

[25] M. Dawgul, M. Maciejewska, M. Jaskiewicz, A. Karafova, and W. Kamysz, "Antimicrobial peptides as potential tool to fight bacterial biofilm.," *Acta poloniae pharmaceutica*, vol. 71,1, pp. 39–47, 2014.

[26] X. Kang, F. Dong, C. Shi, S. Liu, J. Sun, J. Chen, H. Li, H. Xu, X. Lao, and H. Zheng, "DRAMP 2.0, an updated data repository of antimicrobial peptides.," *Scientific Data*, vol. 6(1), 2019.

[27] A. Rajput, A. K. Gupta, and M. Kumar, "Prediction and analysis of quorum sensing peptides based on sequence features," *PLOS ONE*, vol. 10, pp. 1–16, 03 2015.

[28] J. Theolier, I. Fliss, J. Jean, and R. Hammami, "MilkAMP: a comprehensive database of antimicrobial peptides of dairy origin.," *Dairy Science & Technology*, vol. 94(2), p. 181–193, 2013.

[29] K. Nordhausen, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition by Trevor Hastie, Robert Tibshirani, Jerome Friedman.," *International Statistical Review*, vol. 77(3), 2013.

[30] C. D. Fjell, H. Jenssen, K. Hilpert, W. A. Cheung, N. Panté, R. E. W. Hancock, and A. Cherkasov, "Identification of novel antibacterial peptides by chemoinformatics and machine learning," *Journal of Medicinal Chemistry*, vol. 52, no. 7, pp. 2006–2015, 2009. PMID: 19296598.

[31] P. Wang, L. Hu, G. Liu, N. Jiang, X. Chen, J. Xu, W. Zheng, L. Li, M. Tan, Z. Chen, H. Song, Y.-D. Cai, and K.-C. Chou, "Prediction of antimicrobial peptides based on sequence alignment and feature selection methods," *PLOS ONE*, vol. 6, pp. 1–9, 04 2011.

[32] X. Xiao, P. Wang, W.-Z. Lin, J.-H. Jia, and K.-C. Chou, "iamp-2l: A two-level multi-label classifier for identifying antimicrobial peptides and their functional types," *Analytical Biochemistry*, vol. 436, no. 2, pp. 168 – 177, 2013.

[33] P. Schneider, A. T. Müller, G. Gabernet, A. L. Button, G. Posselt, S. Wessler, J. A. Hiss, and G. Schneider, "Hybrid Network Model for "Deep Learning" of Chemical Data: Application to Antimicrobial Peptides.," *Molecular Informatics*, vol. 36, 2016.

[34] S. Liu, J. Bao, X. Lao, and H. Zheng, "Novel 3D Structure Based Model for Activity Prediction and Design of Antimicrobial Peptides.," *Scientific Reports*, vol. 8(1), 2018.

[35] S. A. K. Ong, H. Lin, Y. Chen, Z. Li, and Z. Cao, "Efficacy of different protein descriptors in predicting protein functional families.," *BMC Bioinformatics*, vol. 8(1), 2007.

[36] D. Cao, "propy3. pypi. https://pypi.org/project/propy3/," 2020.

[37] P. J. A. Cock, T. Antao, J. T. Chang, B. A. Chapman, C. J. Cox, A. Dalke, I. Friedberg, T. Hamelryck, F. Kauff, B. Wilczynski, and M. J. L. de Hoon, "Biopython: freely available Python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, pp. 1422–1423, 03 2009.

[38] C. Vens, M.-N. Rosso, and E. G. J. Danchin, "Identifying discriminative classification-based motifs in biological sequences," *Bioinformatics*, vol. 27, pp. 1231–1238, 03 2011.

[39] F. Pedregosa, "Scikit-learn: Machine learning in python. scikit-learn: Machine learning in python.," 2011.

[40] X. Y. Ng, B. A. Rosdi, and S. Shahrudin, "Prediction of Antimicrobial Peptides Based on Sequence Alignment and Support Vector Machine-Pairwise Algorithm Utilizing LZ-Complexity..," *BioMed Research International*, vol. 2015, no. 212715, 2015.

[41] B. Manavalan, T. H. Shin, M. O. Kim, and G. Lee, "Aippred: Sequence-based prediction of anti-inflammatory peptides using random forest," *Frontiers in Pharmacology*, vol. 9, p. 276, 2018.

[42] L. Wang, D. Niu, X. Wang, Q. Shen, and Y. Xue, "A novel machine learning strategy for prediction of antihypertensive peptides derived from food with high efficiency," *bioRxiv*, 2020.

[43] N. W. Schmidt and G. C. L. Wong, "Antimicrobial peptides and induced membrane curvature: geometry, coordination chemistry, and molecular engineering.," *Curr Opin Solid State Mater Sci.*, vol. 17(4), p. 151–163, 2013.

[44] C. Nagant, B. Pitts, K. Nazmi, M. Vandenbranden, J. G. Bolscher, P. S. Stewart, and J.-P. Dehaye, "Identification of peptides derived from the human antimicrobial peptide ll-37 active against biofilms formed by pseudomonas aeruginosa using a library of truncated fragments," *Antimicrobial Agents and Chemotherapy*, vol. 56, no. 11, pp. 5698–5708, 2012.

[45] Z. Gong, X. Pei, S. Ren, X. Chen, L. Wang, C. Ma, X. Xi, T. Chen, C. Shaw, and M. Zhou, "Identification and Rational Design of a Novel Antibacterial Peptide Dermaseptin-AC from the Skin Secretion of the Red-Eyed Tree Frog Agalychnis callidryas.," *Antibiotics.*, vol. 9(5), 2020.

[46] P. Agrawal, D. Bhagat, M. Mahalwal, N. Sharma, and G. P. S. Raghava, "Anticp 2.0: An updated model for predicting anticancer peptides," *bioRxiv*, 2020.

[47] Spengler, Kincses, Mosolygó, Marć, Nové, Gajdács, Sanmartín, McNeil, Blair, and Domínguez-Álvarez., "Antiviral, Antimicrobial and Antibiofilm Activity of Selenoesters and Selenoanhydrides.," *Molecules.*, vol. 24(23), 2019.

[48] J. A. Melvin, L. P. Lashua, M. R. Kiedrowski, G. Yang, B. Deslouches, R. C. Montelaro, and J. M. Bomberger, "Simultaneous antibiofilm and antiviral activities of an engineered antimicrobial peptide during virus-bacterium coinfection," *mSphere*, vol. 1, no. 3, 2016.

[49] A. Friedlander, S. Nir, M. Reches, and M. Shemesh, "Preventing biofilm formation by dairy-associated bacteria using peptide-coated surfaces," *Frontiers in Microbiology*, vol. 10, p. 1405, 2019.

Appendix A

PERFORMANCE

A.1 Model A Performance with Different Features

We evaluated the performance of our main dataset using the SVM algorithm with a different combination of features. We combined features like AAC with motif, DPC with motif, and CTD with the motif. The motif feature was added as per our implemented approach. The performance of the validation dataset is given in the Table 11. All performance metrics are listed in percentage (%) format.

Table 11
Performance Evaluation of Different Features with Model A

Model Performance	**Specificity**	**Sensitivity**	**Accuracy**	**F1 Score**	**MCC**
AAC + Motif	100	79.12	98.16	88.37	88.07
DPC + Motif	99.38	81.25	97.78	86.66	85.67
CTD + Motif	98.96	87.5	97.93	88.41	87.29

A.2 Model B Performance with Different Features

We evaluated the performance of our alternative dataset using the SVM algorithm with a different combination of features. We assessed our model with the SVM as we achieved best performance over other algorithm. We combined features like AAC with motif, DPC with motif, and CTD with the motif.The performance of the validation dataset is given in the Table 12. All performance metrics are listed in percentage (%) format.

A.3 Performance of Regression Model

The original and predicted values of the XGB Regressor are plotted in the Fig. 13. The XGBRegreesor has a higher RMSE and standard deviation than SVR, as observed in the scattered plot.

Table 12
Performance Evaluation of Different Features with Model B

Model Performance	**Specificity**	**Sensitivity**	**Accuracy**	**F1 Score**	**MCC**
AAC + Motif	100	70.83	97.37	82.92	82.97
DPC + Motif	98.76	79.16	96.99	82.60	81.07
CTD + Motif	98.55	85.41	97.37	85.41	83.97

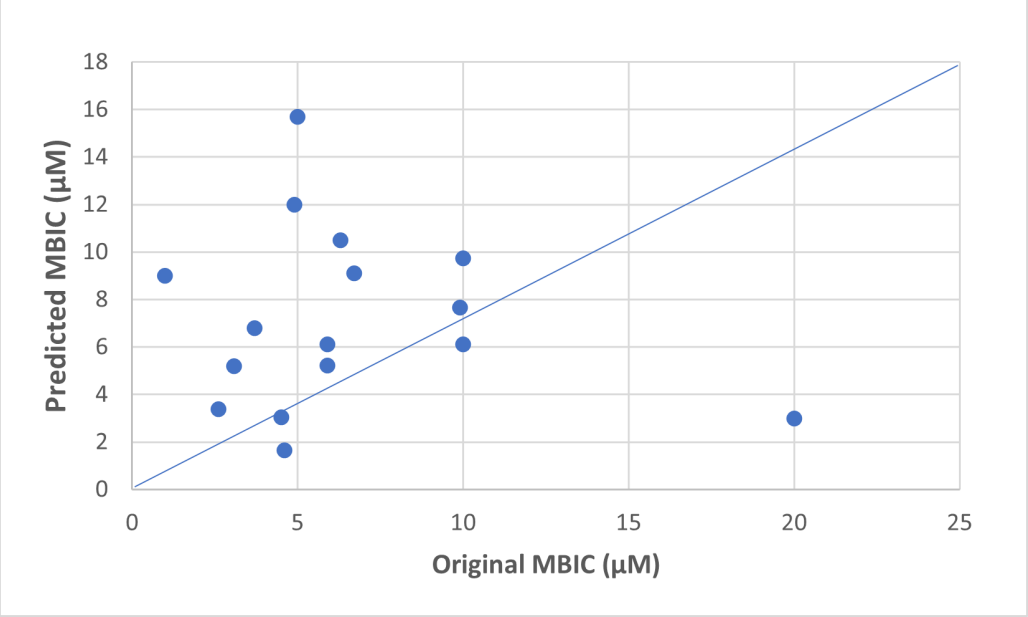

Fig. 13. Distribution of predicted and original MBIC Value from XGBR model.

Appendix B

DATASET

The details of positive dataset with the peptide sequence and length are given in below Fig.(14, 15, 16, 17).

Name	Sequence	Sequence Length
Peptide1	FLGAVLKVAGKLVPAAICKISKKC	24
Peptide2	SLWGKLKEMAAAAGKAALNAVNGLVNQ	27
Peptide3	GFGCPNDYSCSNHCRDSIGCRGGYCKYHVICTCYGCKKRRSIQE	44
Peptide4	FIQHLIPLIPHAIQGIKDIF	20
Peptide5	INWLKLGKAIIDAL	14
Peptide6	GRFKRFRKKFKKLFKKLSPVIPLLHLG	27
Peptide7	GGLRSLGRKILRAWKKYGPIIVPIIRIG	28
Peptide8	SNFDCCLGYTDRILHPKFIVGFTRQLANEGCDINAIIFHTKKKLSVCAN PKQTWVKYIVRLLSKKVKNM	69
Peptide9	RFGRFLRKIRRFRPKVTITIQGSARFG	27
Peptide10	GLFDVIKKVASVIGGL	16
Peptide11	KTKKKLLKKT	10
Peptide12	FWSFLVKAASKILPSLIGGGDDNKSSS	27
Peptide13	VTCDVLSFEAKGIAVNHSACALHCIALRKKGGSCQNGVCVCRN	43
Peptide14	TFPKCAPTRPPGPKPCDINNFKSKFWHIWRA	31
Peptide15	ALWKEVLKNAGKAALNEINNLV	22
Peptide16	GLWSKIKDAAKTAGKAALGFVNEMV	25
Peptide17	KRLFKKLLFSLRKY	14
Peptide18	LGSCVANKIKDEFFAMISISAIVKAAQKKAWKELAVTVLRFAKANGL KTNAIIVAGQLALWAVQCGLS	68
Peptide19	GIFSKLAGKKIKNLLISGLKG	21
Peptide20	GKIIKLKASLKLL	13
Peptide21	VKLFPVKLFP	10
Peptide22	KWAVRIIRKFIKGFIS	16
Peptide23	GIINTLQKYYCRVRGGRCAVLSCLPKEEQIGKCSTRGRKCCRRKK	45
Peptide24	NKGCSACAIGAACLADGPIPDFEVAGITGTFGIAS	35
Peptide25	ILPWKWPWWPWRR	13
Peptide26	FIVPSIFLLKKAFCIALKKC	20
Peptide27	LLGDFFRKSKEKIGKEFKRIVQRIKDFLRNLVPRTES	37
Peptide28	LNLKALLAVAKKIL	14
Peptide29	FFRNLWKGAKAAFRAGHAAWRA	22
Peptide30	GIHDILKYGKPS	12
Peptide31	GLLSGILGAGKKIVF	15
Peptide32	ITSISLCTPGCKTGALMGCNMKTATCHCSIHVSK	34
Peptide33	EVASFDKSKLK	11
Peptide34	FLSLIPHIVSGVASIAKHF	19
Peptide35	FLSMIPKIAGGIASLVKNL	19
Peptide36	FLSLIPAAISAVSALANHF	19
Peptide37	GWGSFFKKAAHVGKHVGKAALTHYL	25
Peptide38	INWLKLGKMVIDAL	14
Peptide39	KTKKKFLKKT	10
Peptide40	RGGRLCYCRRRFCVCVGR	18
Peptide41	KFFKKLKKSVKKHVKKFFKKPKVIGVSIPF	30

Fig. 14. Peptide list of the postive dataset (set1).

Peptide42	LKRVWKRVFKLLKRYWRQLKKPVR	24
Peptide43	RGLRRLGRKIAHGVKKYGPTVLRIIRIAG	29
Peptide44	KWCFRVCYRGICYRKCR	17
Peptide45	FLPFLKSILGKIL	13
Peptide46	LLPIVGNLLKSLL	13
Peptide47	ILPILSLIGGLLGK	14
Peptide48	FLQHIIGALTHIF	13
Peptide49	FLQHIIGALSHFF	13
Peptide50	FFGSVLKLIPKIL	13
Peptide51	WWWLRKIW	8
Peptide52	FIGMIPGLIGGLISAFK	17
Peptide53	FFGTLFKLGSKLIPGVMKLFSKKKER	26
Peptide54	FLGMIPGLIGGLISAFK	17
Peptide55	ILSAIWSGIKSLF	13
Peptide56	VLLVTLTRLHQRGVIYRKWRHFSGRKYR	28
Peptide57	RTCQSQSHRFRGPCLRRSNCANVCRTEGFPGGRCRGFRRRCFCTTHC	47
Peptide58	GRFKRFRKKFKKLFKKLSPVIPLLHL	26
Peptide59	GGLRSLGRKILRAWKKYGPIIVPIIRI	27
Peptide60	RGLRRLGRKIAHGVKKYGPTVLRIIRIA	28
Peptide61	KKVVFKVKFK	10
Peptide62	RWGRWLRKIRRWRPK	15
Peptide63	LLGDFFRKSKEKIGKEFKRIVQRIKDFLRNL	31
Peptide64	RKSKEKIGKEFKRIVQRIKDFLRNL	25
Peptide65	IGKEFKRIVQRIKDFLRNLVPRTES	25
Peptide66	RKSKEKIGKEFKRIVQRIKDFLRNLVPRTES	31
Peptide67	LLGDFFRKSKEKIGKEFKR	19
Peptide68	LLGDFFRKSKEKIGKEFKRIVQRIK	25
Peptide69	LLGDFFRKSKEKI	13
Peptide70	IGKEFKRIVQRIKDFLRNL	19
Peptide71	RKSKEKIGKEFKRIVQRIK	19
Peptide72	IGKEFKRIVQRIK	13
Peptide73	RIVQRIKDFLRNLVPRTES	19
Peptide74	KKVVFWVKFK	10
Peptide75	KSKEKIGKEFKRIVQRIKDFLRNLVPRTES	30
Peptide76	KRIVQRIKDFLRNLVPRTES	20
Peptide77	KRIVQRIKDFLR	12
Peptide78	FKCRRWQWRMKKLG	14
Peptide79	WKLLSKAQEKFGKNKSR	17
Peptide80	RKSYKCLHKRCR	12
Peptide81	KKHRKHRKHRKHGGSGGSKNLRRIIRKGIHIIKKYG	36
Peptide82	RKSYKALHKRAR	12
Peptide83	LAHQKPFIRKSYKCLHKRCR	20
Peptide84	AKRHHGYKRKFH	12

Fig. 15. Peptide list of the postive dataset (set2).

Peptide85	GIGKFLHSAKKFGKAFVGEIMNS	23
Peptide86	KIFGAIWPLALGALKNLIK	19
Peptide87	FHFFHHFFHFFHHF	14
Peptide88	SGSLSTFFRLFNRSFTQALGK	21
Peptide89	TFFRLFNRSFTQALGK	16
Peptide90	KNLRIIRKGIHIIKKY	16
Peptide91	TFFRLFNRSFTQALGKGGGKNLRIIRKGIHIIKKY	35
Peptide92	TFFRLFNRGGGKNLRIIRKGIHIIKKY	27
Peptide93	FKKFWKWFRRF	11
Peptide94	TRRRLFNRSFTQALGKSGGGFKKFWKWFRRF	31
Peptide95	TFFRLFNRSGGGFKKFWKWFRRF	23
Peptide96	KWKLFKKIGAVLKVL	15
Peptide97	KRFRIRVRVIRK	12
Peptide98	RLARIVVIRVAR	12
Peptide99	VQWRIRVRVIKK	12
Peptide100	KQFRIRVRV	9
Peptide101	VQFRIRVRIVIRK	13
Peptide102	KRFRIRVRV	9
Peptide103	VQLRIRVAVIRA	12
Peptide104	VQRWLIVWRIRK	12
Peptide105	IVWKIKRWWVGR	12
Peptide106	RFWKVRVKYIRF	12
Peptide107	RIKWIVRFR	9
Peptide108	VRLRIRVAVRRA	12
Peptide109	IRWRIRVWVRRI	12
Peptide110	RRWVVWRIVQRR	12
Peptide111	IFWRRIVIVKKF	12
Peptide112	VRLRIRVA	8
Peptide113	LRIRWIFKR	9
Peptide114	VRLRIRVAVIRK	12
Peptide115	VRLRIRWWVLRK	12
Peptide116	KRFRIRVAVRRA	12
Peptide117	KRWRWIVRNIRR	12
Peptide118	WRWRVRVWR	9
Peptide119	TFFRLFNRGGGWGSFFKKAAHVGKL	25
Peptide120	RIWVIWRR	8
Peptide121	FLGALFKALSKLL	13
Peptide122	HLGHHALDHLLK	12
Peptide123	ASHLGHHALDHLLK	14
Peptide124	LMCTHPLDCSN	11
Peptide125	VTCDVLSFEAKGIAVNH	17
Peptide126	DSHAKRHHGYKRKFHEKHHSHRGY	24
Peptide127	AKRHHGYKRKFHGGG	15
Peptide128	KKKKKKAAFAAWAAFAA	17

Fig. 16. Peptide list of the postive dataset (set3).

Peptide129	KKKKKKKKKKAAFAAWAAFAA	21
Peptide130	CWFWKWWRRRRR	12
Peptide131	FFGWLIKGAIHAGKAIHGLIHRRRH	25
Peptide132	RWKRWWRRKK	10
Peptide133	RKKRWWRRKK	10
Peptide134	IRIKIRIK	8
Peptide135	IRVKIRVKIRVK	12
Peptide136	DCYCRIPACIAGERRYGTCIYQGRLWAFCC	30
Peptide137	DHYNCVSSGGQCLYSACPIFTKIQGTCYRGKAKCCK	36
Peptide138	GIGKFLHSAGKFGKAFVGEIMKS	23
Peptide139	YSPWTNF	7
Peptide140	ALWKTLLKKVLKA	13
Peptide141	ALWKTLLKKVLKAYSPWTNF	20
Peptide142	RWRWRWF	7
Peptide143	FIKHFIHRFGGGRWRWRWF	19
Peptide144	FIKHFIHRFSATRWRWRWF	19
Peptide145	FIKHFIHRFGGGFKKFWKWFRRF	23
Peptide146	GWKKWLRKGAKHLGQAAIK	19
Peptide147	GQIINLK	7
Peptide148	RWRW	4
Peptide149	RWRWRW	6
Peptide150	RWRWRWRW	8
Peptide151	VNWKKILGKIIKVVK	15
Peptide152	GIGAVLKVLTTGLPALISWIKRKRQQ	26
Peptide153	TLISWIKNKRKQRPRVSRRRRRRGGRRRR	29
Peptide154	CTLISWIKNKRKQRPRVSRRRRRRGGRRRR	30
Peptide155	TLISWIKNKRKQRPRVSRRRRRRGGRRRRC	30
Peptide156	TLISWIKNKRKQCRPRVSRRRRRRGGRRRR	30
Peptide157	LWKTLLKKVLKAAA	14
Peptide158	NEEGFFSARGHRPLDGGGKKKKKK	24
Peptide159	ICIFCCGCCHRSKCGMCCKT	20
Peptide160	KRFKKFFKKLKNSVKKRAKKFFKKPKVIGVTFPF	34
Peptide161	KRFKKFFKKLKNSVKKRFKKFFKKLKVIGVTFPF	34
Peptide162	FKCRRWQWRMKKLGAPSITCVRRAF	25
Peptide163	GLKLRFEFSKIKGEFLKTPEVRFRDIKLKDNRISVQR	37
Peptide164	RFRRLFRIRVRVLKKI	16
Peptide165	FRIRVRV	7
Peptide166	AFKAFWKFVKFVK	13
Peptide167	KWFWKFVKFVK	11
Peptide168	IKKILSKIKKLLK	13
Peptide169	GRRRRSVQWCA	11
Peptide170	YAPWTNF	7
Peptide171	KRWWKWWRRC	10
Peptide172	IRWRIRVWVRRIC	13

Fig. 17. Peptide list of the postive dataset (set4).

Peptide173	KRWRIRVRVIRKC	13
Peptide174	WIVVIWRRKRRRC	13
Peptide175	YAPWTNA	7
Peptide176	RILSILRHQNLLKELQDLAL	20
Peptide177	KWKVFKKIEKMGRNIRNGIVKAGPAIAVLGEAKAL	35
Peptide178	GIGLFLHSAGLFGLAFVGEIMKS	23
Peptide179	CVNWKKILGKIIKVVK	16
Peptide180	FFGKVLKLIRKIF	13
Peptide181	GRWKRWRKKWKKLWKKLS	18
Peptide182	RLCRIVVIRVCR	12
Peptide183	KWKLFKKIGIGKFLHSAKKF	20
Peptide184	RPAFRKAAFRVMRACV	16
Peptide185	LLLFLLKKRKKRKY	14
Peptide186	GIWKKWIKKWLKKLLKKLWKKG	22
Peptide187	LAREYKKIVEKLKRWLRQVLRTLR	24
Peptide188	IGKEFKRIVERIKRFLRELVRPLR	24
Peptide189	MLCVLQGLRE	10
Peptide190	ELRLVCMGQL	10
Peptide191	MLCVLQGLREGG	12
Peptide192	MLCVLQGLREC	11
Peptide193	VRLIVAVRIWRR	12
Peptide194	RRWIRVAVILRV	12
Peptide195	VRLIRAVRAWRV	12
Peptide196	VRWARVARILRV	12
Peptide197	VRLIWAVRIWRR	12
Peptide198	VRLIVRIWRR	10
Peptide199	RFKRVARVIW	10
Peptide200	IGIKLLKSKLKAL	13
Peptide201	IKIKIKIK	8
Peptide202	FKKVIVIRRWFI	12
Peptide203	KRIRWVILWRQV	12
Peptide204	VFLRRIRVIVIR	12
Peptide205	RIVIVRIRRLFV	12
Peptide206	VFWRRIRVWVIR	12
Peptide207	RIVWVRIRRWFV	12
Peptide208	VQLRAIRVRVIR	12
Peptide209	RIVRVRIARLQV	12
Peptide210	VQLRRIRVWVIR	12
Peptide211	RIVWVRIRRLQV	12
Peptide212	VQWRAIRVRVIR	12
Peptide213	RIVRVRAIRWQV	12
Peptide214	VQWRRIRVWVIR	12
Peptide215	RIVWVRIRRWQV	12
Peptide216	AKRRRGYKRKFKK	13

Fig. 18. Peptide list of the postive dataset (set5).

Peptide217	GTPGPQGIAGQRGVV	15
Peptide218	GTPGPQGIAGQRGVVAEAAAKEAAAKEAAAKASGSLSTFFRLFNRS FTQALGK	53
Peptide219	CGGLLLFLLKKRKKRKY	17
Peptide220	NGVQPKYKWWKWWKKWW	17
Peptide221	NGVQPKYRWWRWWRRWW	17
Peptide222	PFWRIRIRR	9
Peptide223	PFFWRIRIRR	10
Peptide224	FWRRFWRR	8
Peptide225	FWRIRIRR	8
Peptide226	KRAKKFFKKLK	11
Peptide227	KRAKKFFKKPK	11
Peptide228	KRFKKFFKKLK	11
Peptide229	LKLLKKLLKKLLKLL	15
Peptide230	KKKLLLLLLLLLKKK	15
Peptide231	LLLLLKKKKKLLLL	15
Peptide232	KNLRRIIRKGIHIIKKYG	18
Peptide233	WKKIRVRLSA	10
Peptide234	KWKIRVRLSA	10
Peptide235	KIKWILKYWKWS	12
Peptide236	RIRWILRYWRWS	12
Peptide237	GLLWHLLHHLLH	12
Peptide238	LAAKLTKAATKLTAALTKLAAALT	24
Peptide239	DGVKLCDVPSGTWSGHCGSSSKCSQQCKDREHFAYGGACHYQFPS VKCFCKRQC	54
Peptide240	EHFAYGGAKHYQFPSVKKFKKRQK	24
Peptide241	RRRWWWWV	8
Peptide242	FLSLIPKIAGGIAALAKHL	19

Fig. 19. Peptide list of the postive dataset (set6).