Chapter 1

Introduction

1.1 Introduction

In 1900s, infectious diseases were a leading cause of death. However, in a few decades, antibiotics have drastically changed modern medicine and extended the average human lifespan by 23 years. Antibiotic consumption has continued to increase unchecked among low and middle-income countries. As a result, the microbes have evolved rapidly and developed resistance to conventional antimicrobial drugs. Due to the developed resistance, antimicrobial drugs become ineffective against the microbes. This phenomenon is also known as antimicrobial resistance (AMR). AMR reportedly kills at least seven lakh people every year. Due to its severe impact on humans, the world health organization (WHO) has declared it as one of the top ten global public health threats. Thus, the need of the hour is to discover novel antimicrobial compounds for combating this alarming situation. Among various approaches for discovering novel antimicrobial compounds, antimicrobial peptides (AMPs) seem promising. AMPs are proteins produced by a diverse population of living organisms, and depending upon their target microbe, AMPs can be classified into multiple groups, such as antiviral peptides (AVPs), antifungal peptides (AFPs), antibacterial peptides (ABPs), etc.

Identifying novel AMPs in the lab by performing different experiments involves a

lot of costs and also hampers the timely discovery of peptide-based drugs. Therefore wet lab researchers utilize in-silico tool(s) for preliminary screening of natural sources for identifying potential AMPs. The existing tools available to serve this purpose have poor generalization performance, which limit their applicability for wet-lab researchers. Thus, we have developed AI-based tools for identifying AFPs, AVPs, and ABPs from natural sources in different studies.

Peptides are made up of amino acids. All the amino acids are not equally important for classifying the peptide into a particular class. The state-of-the-art AI-based frameworks suffer from the limitation of being black boxes. As a result, they cannot provide information about the amino acids that play an essential role in the classification of a peptide (known as critical amino acids). For better understanding, we can relate the task of classifying the peptide to the sentiment analysis task, where given a sentence, the goal is to determine whether expressed opinion in the sentence is positive or negative. In the case of sentiment analysis, we can see that not all words contribute equally in determining the sentiment of a sentence (some words have more contribution, some words have less contribution, and others have negative contribution). Consider the statement, "This film is not good". The overall sentiment of this statement is negative. The word that contributed most towards the prediction is "not", while the word "good" contributed against the prediction, and the remaining words contributed little or nothing. In the sentiment analysis task, identifying critical words is possible for normal people, but identifying critical amino acids in the case of peptides is not at all possible without thorough lab-based experimentation and evaluation (which is time consuming and labor-intensive). Therefore, there is a strong need for explainability in the task of peptide classification. Thus, in this study, we proposed a framework that provides information about the amino acids that play an essential role in classifying a peptide.

The WHO categorizes bacteria into three categories of priority: critical, high, and

medium, according to the urgency of the need to develop new antibiotics to combat these pathogens. Amongst the drug-resistant bacteria, the ESKAPEE (*Enterococcus faecium, Staphylococcus aureus, Klebsiella pneumoniae, Acinetobacter baumannii, Pseudomonas aeruginosa, Enterobacter spp., and Escherichia coli*) pathogens pose a major threat, as these range from high to critical WHO-priority pathogens. The main drawback of earlier studies is that they do not provide the minimum inhibitory concentration (MIC) values against the ESKAPEE pathogens for the identified ABP. Therefore, after identifying ABPs, wet-lab researchers have to test them against the ESKAPEE pathogens at different concentrations, leading to a loss of time and money. Also, since it involves human intervention, there might be a case where we may lose optimal ABP(s), which can work at low MIC(s) against the entire ESKAPEE group of bacteria. To address this, we conducted a study and proposed a framework that predicts the MIC values for the ABP against the ESKAPEE pathogens.

Although much research has been undertaken to identify peptide-based drugs, only a few are commercially accessible. The toxicity of peptides is the key roadblock to the development of therapeutic peptides as medications. Numerous laboratory-based techniques for determining peptide toxicity are available, out of which the hemolytic activity of peptides against red blood cells (RBCs) is regarded as the first-line technique. Hemolysis is a condition when RBCs are destroyed before reaching their anticipated lifetime of 120 days. The situation becomes more severe when the destruction of RBCs exceeds its creation due to hemolysis. As a result, the hemoglobin level and oxygen-carrying capability of blood decrease, leading to anemia. Therefore, highly hemolytic peptides are not suitable for pharmaceutical formulations. As a result, identifying low hemolytic peptides is crucial for creating novel peptide-based therapeutics. Experiments to discover low hemolytic peptides are labor-intensive, time-consuming, and involve testing mammalian red blood cells. To address this, in this study, we proposed a framework that will help in identifying low hemolytic therapeutic peptides.

The proposed frameworks from the aforementioned studies are made available as web-based tools. The wet-lab researchers can use these tools to find low hemolytic AMPs effective against different pathogens.

1.2 Preliminaries

The preliminaries of this dissertation are categorized into two groups; biological aspects and computational aspects. The following is a description of these two aspects.

1.2.1 Biological aspects

This section describes the basics of antimicrobial resistance (AMR) snd antimicrobial peptides (AMPs).

1.2.1.1 Antimicrobial resistance (AMR)

In 1900s, infectious diseases were a leading cause of death. Later, with the discovery of antimicrobial drugs the average human lifespan got extended by 23 years (Figure 1.1). But, due to the overuse of antimicrobial drugs, pathogen species perform different mutations in their genome, which help them to develop resistance against these drugs (also known as antimicrobial resistance (AMR)). As a result of AMR pathogens start degrading antimicrobial medications, which also ensures their survival in the presence of antimicrobial drugs (Figure 1.1). As a result WHO has declared AMR as one of the top 10 global public health threats. Additionally, AMR does not confine itself to the source, but rather spreads quickly across a variety of channels, as illustrated in Figure 1.2

1.2.1.2 Antimicrobial peptides (AMPs)

- Amino Acids : Amino acids are organic compounds that contain both amino and carboxylic acid functional groups. Each amino acid has at least one amino (NH_2)

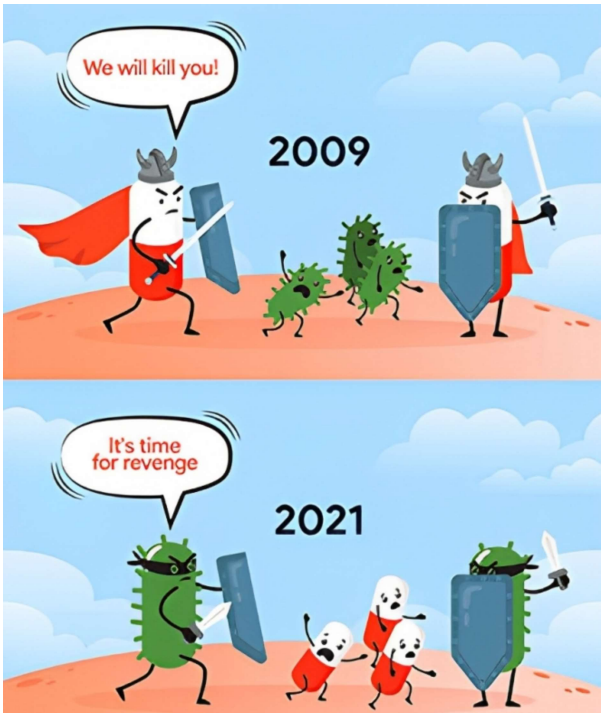

Figure 1.1: Antimicrobial Resistance (Source:[1])

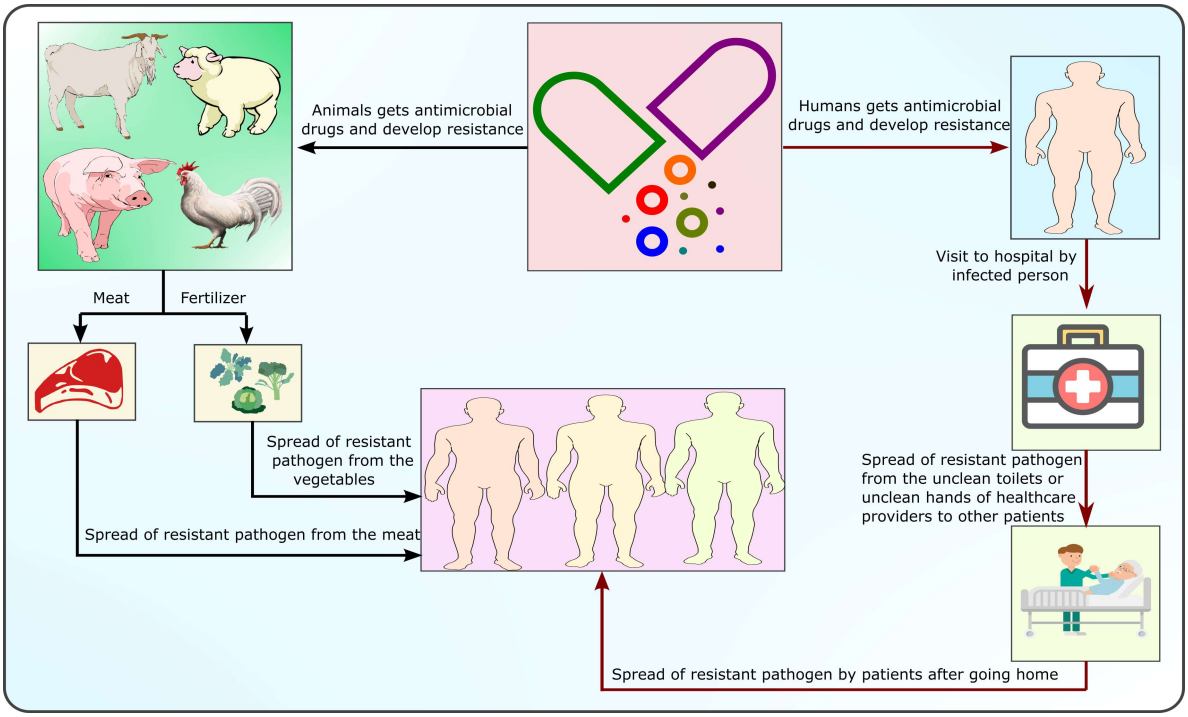

Figure 1.2: Spread of Antimicrobial resistance

and a carboxylic acid group (COOH). For different value of R at alpha carbon we get different amino acid.

- Peptide: A peptide is formed when amino acids are joined together using peptide bond. The water molecule got released during this process.
- AMPs: These are small peptides that play a crucial role in the innate immune system of various organisms. These peptides possess broad-spectrum antimicrobial activity and can inhibit the growth of bacteria, fungi, parasites, and viruses.

1.2.2 Computational Aspects

This section describes distinct computational aspects which forms the core of frameworks proposed as part of different studies:

1.2.2.1 Gated Recurring Units

The GRU is made up of a series of cells that repeat themselves. At each time step t ($1 \leq$ t ≤ 50=(maximum sequence length)), cell takes x_t and h_{t-1} as input and outputs h_t. The following equations describe the computational steps involved in each cell:

$$z_t = \sigma(W_z[h_{t-1}, x_t] + b_z) \tag{1.1}$$

$$r_t = \sigma(W_r[h_{t-1}, x_t] + b_r) \tag{1.2}$$

$$\tilde{h}_t = tanh(W_h[r_t \otimes h_{t-1}, x_t] + b_h) \tag{1.3}$$

$$h_t = (1 - z_t) \otimes h_{t-1} + z_t \otimes \tilde{h}_t \tag{1.4}$$

In the above equations, x_t stands for the embedding vector of the amino acid A_t at a given timestep t. z_t, r_t, $\tilde{h}_t$, represent update gate, reset gate and candidate state, respectively, which helps in computing h_t. (W_z, b_z), (W_r, b_r) and (W_h, b_h) represent the weights and biases of the update gate, reset gate and candidate state, respectively.

$\otimes$ denotes the element-wise multiplication operation.

1.2.2.2 Long short-term memory

The LSTM is made up of a series of cells that repeat themselves. At each time step t ($1 \leq$ t $\leq$ maximum sequence length), cell takes x_t, h_{t-1} and c_{t-1} as input and outputs h_t, c_t . The following equations explains the calculation of h_t and c_t:

$$\textbf{Input gate :}\ i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \tag{1.5}$$

$$\textbf{Output gate :}\ o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \tag{1.6}$$

$$\textbf{Forget gate :}\ f_t = \sigma(W_f[h_{t-1}, x_t] + b_f) \tag{1.7}$$

$$\begin{aligned}\textbf{Cell state :}\ c_t = (f_t \otimes c_{t-1}) \oplus \\ (i_t \otimes tanh(W_c[h_{t-1}, x_t] + b_c))\end{aligned} \tag{1.8}$$

$$\textbf{Hidden state :}\ h_t = o_t \otimes tanh(c_t) \tag{1.9}$$

In the above equations, the embedding vector of the amino acid A_t at a given time step t is represented by x_t. The weights and biases of i_t, o_t, f_t, and c_t are represented by (W_i, b_i), (W_o, b_o), (W_f, b_f), and (W_c, b_c), respectively. The element-wise multiplication and addition operations are denoted by $\otimes$ and $\oplus$, respectively.

1.2.2.3 Temporal Convolutional Network

A TCN [6, 7, 8, 9] is a deep learning architecture made by modifying a Convolutional Neural Network (CNN). Unlike standard convolution operation, TCN uses dilated causal convolution operation. Different concepts related to TCN are given below:

(i) **Causal convolutional**: The convolutional operation performed to obtain the output at time T considers the input from 0 to T.(ii) **Dilations**: One of the issues

with causal convolutions is that in the case of causal convolutions receptive field scales linearly with depth, necessitating the use of multiple layers to increase the receptive field. The solution to this is to use the concept of dilations (skipping values between the inputs of the convolutional operation) with casual convolutions, which increase the receptive field without significantly increasing the computational cost. (iii) **Receptive Field**: The maximum number of steps back in time from the current sample at time T that a filter can hit. Ideally, the receptive field must be bigger than the largest length of the input sequence. If a sequence longer than the receptive field is passed into the model, extra values further back in the sequence will be replaced with zeros. Let $D(i)$ denote dilation corresponding to layer i, $R(i)$ denotes number of time steps dilated causal convolution layer i can see, k denotes the kernel size and n denotes number of dilations then the receptive field $R(n)$ for dilated causal convolutions can be calculated as follows:

$$R(1) = 1 + (k-1)D(1) \tag{1.10}$$

$$R(2) = R(1) + (k-1)D(2) \tag{1.11}$$

$$R(n-1) = R(n-2) + (k-1)D(n-1) \tag{1.12}$$

$$R(n) = R(n-1) + (k-1)D(n) \tag{1.13}$$

Substituting Equations 1.10 - 1.12 in Equation 1.13:

$$\begin{aligned} R(n) = 1 &+ (k-1)D(1) + (k-1)D(2) \\ &+ .... + (k-1)D(n-1) + (k-1)D(n) \end{aligned} \tag{1.14}$$

Equation 1.14 can be written as:

$$R(n) = 1 + (k-1)(D(1) + ... + D(n-1) + D(n)) \tag{1.15}$$

If dilation corresponding to each layer increases exponentially then D(1) = 1, D(2) = 2, D(3) = 4 ,...., D(n-1) = 2^{n-2}, D(n) = 2^{n-1}. Substituting these values in Equation 1.15:

$$R(n) = 1 + (k-1)(1 + 2 + 4 + ... + 2^{n-2} + 2^{n-1}) \tag{1.16}$$

Substituting the value of $1 + 2 + 4 + ... + 2^{n-2} + 2^{n-1}$ as $2^n - 1$ in Equation 1.16:

$$R(n) = 1 + (k-1)(2^n - 1) \tag{1.17}$$

In TCN, residual blocks are used to prevent vanishing and exploding gradient problems. Each residual block contains two identical dilated causal convolutions (same kernel sizes and dilations). The outputs of the block are obtained by adding the results of the final convolution back to the inputs.

1.2.2.4 1D convolutional neural networks

The convolution operation is the core component of 1DCNN. Initially, CNN was designed to discover the spatial patterns in the images. This was accomplished by sliding a small window of size (height x width) across an image in horizontal and vertical directions, followed by performing a convolution operation between the image and the window. When using the CNN with peptides, the convolution procedure requires the window to examine the complete embedding vector corresponding to each amino acid. As a result, the width of the window becomes fixed, and it can now only move in one direction. Thus, the convolution operation is called a 1D convolution operation, and CNN is known as 1DCNN.

1.2.2.5 Transfer learning

Transfer learning helps to learn a new task by transferring knowledge from a related task that has already been learned, which saves time and helps in better generalization

[10, 11, 12]. Transfer learning can be applied to both image and text data. In the case of image data, the concept of transfer learning is usually realized by utilizing the pretrained weights from the model trained on a large image dataset, whereas in the case of text data, the idea of transfer learning is usually realized by adopting the pretrained embeddings from the model trained on a large text dataset [13]. Taking into consideration the advantage of transfer learning, we have utilized it with peptides. The concept of transfer learning was accomplished by utilizing pretrained embeddings from [14] (Authors learned these embeddings by training Embeddings from Language Models (ELMo) on millions of protein sequences from UniRef50) and from [15] (Authors obtained these pretrained embeddings by training the 33-layer transformer model on millions of protein sequences from UniRef50 in an unsupervised manner).

1.2.2.6 Performance Metrics For Classification Models

To evaluate the performance of classification models, various performance metrics namely Accuracy (A_{cc}), Sensitivity(S_n), Precision(P_r), F1-Score (F_s), Specificity (S_p), Matthews correlation coefficient (MCC) are used . These metrics are defined in Equations 1.18 - 1.23

- True Positive (TP): Number of peptides correctly classified as ABPs.
- True Negative (TN): Number of peptides correctly classified as Non-ABPs.
- False Positive (FP): Number of peptides misclassified as ABPs.
- False Negative (FN): Number of peptides misclassified as Non-ABPs.

Accuracy (A_{cc}): Ratio of accurately classified peptides to the total number of peptides.

$$A_{cc} = \frac{TP + TN}{TP + TN + FP + FN} \tag{1.18}$$

Sensitivity(S_n): Correct prediction ratio of positive samples. It is also known as recall or TPR.

$$S_n = \frac{TP}{TP + FN} \tag{1.19}$$

Precision(P_r): Ratio of accurately classified ABPs to the total number of peptides that were classified as ABPs. It is also known as the positive predictive value (PPV).

$$P_r = \frac{TP}{TP + FP} \tag{1.20}$$

F1-Score (F_s): Harmonic mean of S_n and P_r.

$$F_s = \frac{2 \times S_n \times P_r}{S_n + P_r} \tag{1.21}$$

Specificity (S_p): Ratio of accurately classified Non-ABPs to the actual number of Non-ABPs present in the dataset. It is also known as true negative rate (TNR).

$$S_p = \frac{TN}{TN + FP} \tag{1.22}$$

Matthews correlation coefficient (MCC): Provides the correlation between actual and predicted values and is regarded as the best metric to evaluate the performance of the model on an imbalanced dataset.

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}} \tag{1.23}$$

1.2.2.7 Performance Metrics For Regression Models

To evaluate the performance of regression models, various performance metrics namely Pearson correlation coefficient (PCC), coefficient of determination (COD), mean absolute error (MAE), mean squared error (MSE), and root mean squared error (RMSE) were used.

$$PCC = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2(y_i - \overline{y})^2}} \tag{1.24}$$

$$COD = 1 - \frac{\sum_{i=1}^{n}(x_i - y_i)^2}{\sum_{i=1}^{n}(x_i - \overline{x})^2}. \tag{1.25}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - x_i| \tag{1.26}$$

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2 \tag{1.27}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i)^2} \tag{1.28}$$

Where, x_i and y_i denote the actual and predicted value, respectively. $\overline{x}$ and $\overline{y}$ denote the mean of actual and predicted values, respectively.

1.2.2.8 Adam Optimizer

Adam is a stochastic gradient descent method that is based on adaptive estimation of first-order and second-order moments. According to [16], the method is computationally efficient, has little memory requirement, invariant to diagonal rescaling of gradients, and is well suited for problems that are large in terms of data/parameters. In Adam, exponential moving averages m_t and v_t of gradient g_t and square of the gradient g_t^2, respectively, were employed. Both m_t and v_t were initialized as zero vectors, and the value of both m_t and v_t got updated in each iteration t. Adam's parameter update for

each iteration t is as follows:

$$g_t = \frac{\partial L}{\partial w}\Big|_{w_{t-1}} \tag{1.29}$$

$$m_t = \alpha m_{t-1} + (1-\alpha) g_t \tag{1.30}$$

$$v_t = \beta v_{t-1} + (1-\beta) g_t^2 \tag{1.31}$$

$$\hat{m}_t = \frac{m_t}{1-\alpha^t} \tag{1.32}$$

$$\hat{v}_t = \frac{v_t}{1-\beta^t} \tag{1.33}$$

$$w_t = w_{t-1} - \eta\left(\frac{\hat{m}_t}{\sqrt{\hat{v}_t} + \epsilon}\right) \tag{1.34}$$

Where, L is the binary cross-entropy loss function, ϵ is a small number used to prevent division by zero, α and β control the decay rates of m_t and v_t, respectively (both have their default value close to 1), η denotes the learning rate.

1.2.2.9 Activation Functions

Activation function is used to introduce non-linearity into the output of a neuron. This helps the network to learn complex patterns in the data. Different activation functions used are given below:

1. Rectified linear unit (ReLU) activation function:

$$ReLU(x) = max(0, x) = \begin{cases} 0, \text{ if } x \leqslant 0 \\ x, \text{ if } x > 0 \end{cases} \tag{1.35}$$

2. Sigmoid activation function:

$$Sigmoid(x) = \frac{1}{1+e^{-x}} \tag{1.36}$$

3. Softmax activation function:

$$Softmax(z_i) = \frac{e^{z_i}}{\sum_{j=1}^{2} e^{z_j}} \tag{1.37}$$

1.3 Motivation and Objective of the Thesis

AMPs are found in nature, but their wet lab identification is challenging due to the involvement of (i) Time : There are millions of protein sequences, and identifying AMPs from them in the lab using the hit-and-trial method involves a lot of time. (ii) Cost : AI can reduce the drug discovery cost by a large margin (Figure 1.3). A lot of peptides need to be synthesized for conducting different lab experiments using the hit-and-trial method, which involves a lot of cost. Therefore, AI can be used as a funnel having different filtering layers so that the optimal AMPs get selected, which can be then chosen based on domain expertise, synthesised and used to do *in-vitro* and *in-vivo* studies by wet lab researchers (Figure 1.4)

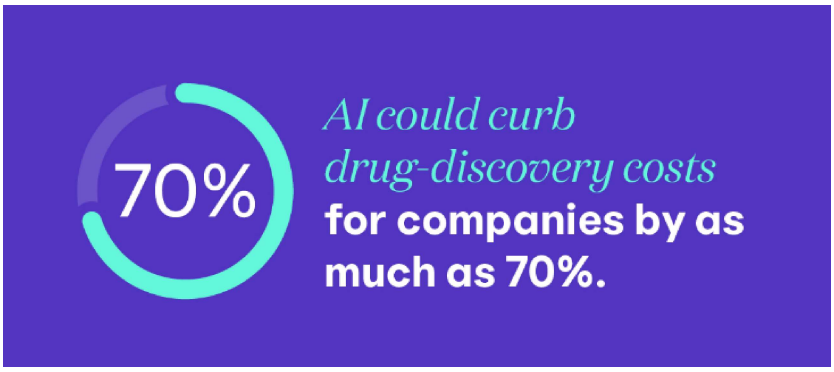

Figure 1.3: AI in Drug Discovery (Source:[2])

Multiple studies in the past have developed different tools (like iAMPpred [17], Antifp [18], PhytoAFP [19], AVPpred [20], Meta-iAVP [21], AVPIden [22], ENNAVIA [23], HEMOPI [5], HemoPred [4], HLPpred-Fuse [3]) to aid wet lab researchers in discovering antimicrobial peptide-based medications to combat the alarming situation of antimicrobial resistance.

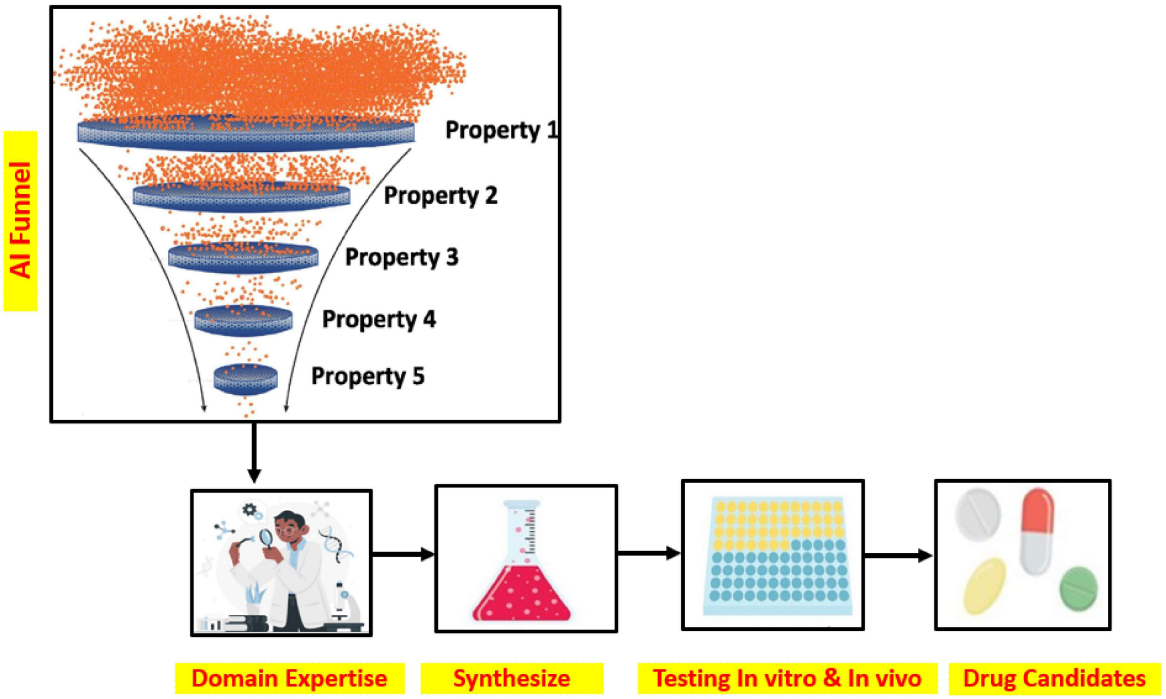

Figure 1.4: AI in peptide based antimicrobial drug discovery.

iAMPpred [17] was developed utilizing SVM machine learning algorithm with the handcrafted features (HCF) for classification of AFPs. These HCF were prepared by considering various properties of peptides like alpha-helix propensity , beta-turn propensity , beta-sheet propensity , charge , hydrophobicity index , isoelectric point , molecular weight , amino acid composition , pseudo amino acid composition, amphiphilic pseudo amino acid composition etc. Antifp [18] was developed utilizing SVM machine learning algorithm with N15C15 binary profile-based features of peptides for classification of AFPs. PhytoAFP [19] was developed utilizing SVM machine learning algorithm with tripeptide composition-based features of peptides for classification of AFPs. AVPpred [20] was developed utilizing SVM machine algorithm with the HCF for classification of AVPs. These HCF were prepared by considering various properties of peptides like amino acid composition, charge, hydrophobicity, secondary structure, charge, size, residue composition, hydrophobicity, etc. Meta-iAVP [21] was developed utilizing meta classifier comprising of six different machine learning algorithms (RF,

SVM, KNN, DT, LR, and XGBoost) with amino acid composition , pseudo amino acid composition based features with the for classification of AVPs. AVPIden [22] was developed utilizing RF machine learning algorithm with HCF for classification of AVPs. These HCF were prepared by considering various compositional properties of peptides like amino acid composition, dipeptide composition, pseudo amino acid composition,etc and various physicochemical features like aliphatic index, alpha helical propensity, transmembrane propensity, hydrophobic moment, hydrophobicity, Boman index, isoelectric point , net charge etc. ENNAVIA [23] was developed utilizing neural networks with HCF constructed by considering various compositional, physicochemical and structural properties of peptides for classification of AVPs. HEMOPI [5] was developed utilizing SVM machine learning algorithm with HCF constructed by considering amino acid composition, dipeptide composition, binary profile properties of peptides for identifying hemolytic activity. HemoPred [4] was developed utilizing RF machine learning algorithm with HCF constructed by considering amino acid composition, dipeptide composition and physicochemical properties of peptides for identifying hemolytic activity. HLPpred-Fuse [3] was developed utilizing amino acid composition, dipeptide composition, amino acid index, binary profile, composition transition distribution, conjoint triad, quasi-sequence order, grouped dipeptide composition, grouped tripeptide composition with meta classifier comprising of six different machine learning algorithms (SVM, RF, GB, ERT, KNN, AB) for identifying hemolytic activity.

The aforementioned tools have the following drawbacks, which limit their applicability for wet-lab researchers (i) They were developed using traditional machine-learning techniques. (ii) Data is the food for Artificial intelligence (AI), and the generalization performance of a model strongly depends on it. As a result, there has been a recent push in the AI community toward data-centric AI from model-centric AI [24, 25, 26, 27]. With the advancement in time, technology, and the need to develop alternatives for traditional antibiotics, the literature on AMPs has expanded significantly. However, the

existing tools have utilized only a few available AMPs in the literature. (iii) Only a few keywords were considered while developing a filter for extracting the negative set from the literature, which may have caused even positive class peptides to cross the filter. (iv) Peptides are made up of amino acids. All the amino acids are not equally important for classifying the peptide into a particular class. The existing frameworks suffer from the limitation of being black boxes. As a result, they cannot provide information about the amino acids that play an essential role in the classification of a peptide (known as critical amino acids). (v) They do not provide the minimum inhibitory concentration (MIC) values against the ESKAPEE pathogens for the identified ABP. Therefore, after identifying ABPs, wet lab researchers have to test them against the ESKAPEE pathogens at different concentrations, leading to a loss of time and money. Apart from this, due to the involvement of time and money, testing all the identified ABPs in the lab at different concentrations is not feasible, which may lead to loss of optimal ABP(s), which can work at low MIC(s) against all the ESKAPEE group of bacteria

Deep learning algorithms can automatically learn the optimal features from the data, thus reducing our reliance on domain experts and, in most cases, outperform machine learning algorithms. The concept of transfer learning and ensemble learning can also be used with deep learning algorithms, which can boost performance. Besides this, a Multimodal AI approach that combines both HCF and DLF can be used to enhance performance further.

The main objective of the thesis is to aid wet lab researchers in the discovery of novel optimal low hemolytic antimicrobial peptide-based medications. In light of this, keeping in mind the shortcomings of the previous research and motivated by the deep learning, transfer learning, ensemble learning, and Multimodal AI approach, we have designed our objectives. The main objectives of this dissertation are as follows:

1. To develop AI-based tools that can identify AFPs and AVPs from natural resources and help fight the recurring viral and fungal outbreaks.

2. To develop an explainable tool that can provide information about the critical amino acids responsible for the prediction. Wet-lab researchers can use this information to make certain decisions based on their domain expertise.
3. To develop a AI assisted tool that can identify optimal ABPs from a large pool of ABPs that can work at low MICs against the WHO-priority ESKAPEE pathogens.
4. To develop an AI-based tool that can help identify low hemolytic peptides and help in developing antimicrobial peptide-based medications that are suitable for pharmaceutical formulations.

1.4 Contributions

The overall layout of the thesis has been depicted in Figure 1.5. In this dissertation, we have five contributing chapters. A brief description about them is as follows:

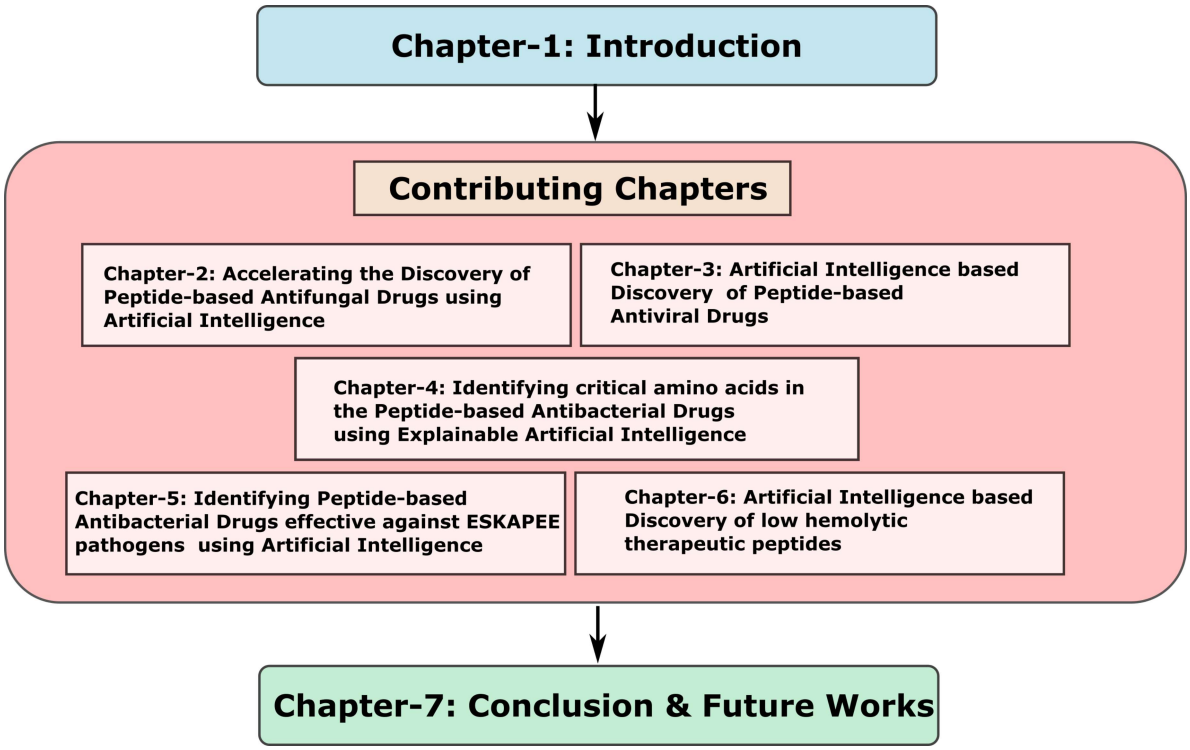

Figure 1.5: Layout of the Thesis

Chapter-2: This chapter introduces a model named Deep-AFPpred, which can identify AFPs from natural resources. This model utilizes the concept of transfer learning with a deep learning algorithm and has better generalization performance than existing *in-silico* tools. The concept of transfer learning was accomplished by utilizing pretrained embeddings from [14]. These pretrained embeddings were learned by training Embeddings from Language Models (ELMo) on millions of protein sequences from UniRef50. Deep-AFPpred is based on deep learning that does not require hand-crafted features (HCF) for making predictions, thus removing our reliance on domain expertise. As a proof of concept, novel AFPs are also identified using Deep-AFPpred in this chapter by screening ten antifungal proteins from five distinct genera (two antifungal proteins from each genus). To assist wet-lab researchers in identifying novel AFPs from any protein, the proposed model Deep-AFPpred has also been made available as a web server at https://afppred.anvil.app/.

Chapter-3: It presents a novel model named Deep-AVPpred, which can identify AFPs from natural resources. This model utilizes the concept of transfer learning with a deep learning algorithm and has better generalization performance than existing *in-silico* tools. The concept of transfer learning was accomplished by utilizing pretrained embeddings from [15]. These pretrained embeddings were learned by training the 33-layer transformer model on millions of protein sequences from UniRef50 in an unsupervised manner. Deep-AVPpred is based on deep learning that does not require HCF for making predictions, thus removing our reliance on domain expertise. As a proof of concept, novel AVPs are also identified using Deep-AVPpred in this chapter from antiviral proteins belonging to the human interferon-α family. The model is also deployed as a web server to assist researchers in discovering novel AVPs from protein sequences and is freely available online at https://deep-avppred.anvil.app/.

Chapter-4:This chapter proposes an explainable artificial intelligence-based framework named XAI-INVENT, which can discover novel peptide antibiotics. This frame-

work is the first of its kind, which not only identifies potent ABPs from protein sequences but also provides information about the amino acids that play an essential role in the classification of a peptide. As a proof of concept, novel ABPs are also identified using XAI-INVENT in this chapter from bacteriocin acquired from the ESKAPEE group of bacteria, which are highly threatening drug-resistant WHO priority I and II pathogens in this chapter. To help researchers find new ABPs from protein sequences, the model has been set up as a web server and is freely accessible online at https://xai-invent.anvil.app/

Chapter-5: Here we propose a model named ESKAPEE-MICpred, which supplements the currently available ABP classification tools and provides MIC values against the ESKAPEE pathogens for a given ABP. As proof of concept, this chapter identifies five ABPs from the therapeutic peptides and five novel ABPs from the antibacterial protein sequences using ESKAPEE-MICpred. The proposed model has been deployed as a web server at https://eskapee-micpred.anvil.app/ to aid the scientific community.

Chapter-6: Here we propose a model named EnDL-HemoLyt, which can identify low hemolytic peptides. This model integrates deep learning algorithms using the min/max combiner ensemble technique and uses features from DLF and HCF. The proposed model is trained and tested on the recent dataset, which contains standard peptides and peptides with either N-terminal acetylation or C-terminal amidation, or both. Ablation studies have also been performed to understand the contribution of the ensemble algorithm, HCF, and DLF. To help wet-lab researchers identify low hemolytic therapeutic peptides, the model developed from the proposed framework has been set up as a web server and is freely accessible at https://endl-hemolyt.anvil.app/.