ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

DEEP LEARNING FOR ANTIMICROBIAL PEPTIDE RECOGNITION

HÜSEYİN OKAN SOYKAM

M.Sc. THESIS

DEPARTMENT OF BIOSTATISTICS AND BIOINFORMATICS

SUPERVISOR
Assoc. Prof. Günseli Bayram Akçapınar

ISTANBUL-2022

ACIBADEM MEHMET ALI AYDINLAR UNIVERSITY
INSTITUTE OF HEALTH SCIENCES

DEEP LEARNING FOR ANTIMICROBIAL PEPTIDE RECOGNITION

HÜSEYİN OKAN SOYKAM

M.Sc. THESIS

DEPARTMENT OF BIOSTATISTICS AND BIOINFORMATICS

SUPERVISOR
Assoc. Prof. Günseli Bayram Akçapınar

ISTANBUL-2022

DECLARATION

I declare that this thesis work is my own work, I had no unethical behavior at any stages from the planning to the writing of the thesis, I obtained all the information in this thesis in accordance with academic and ethical rules, I cited all the information and comments that were not obtained with this thesis work, and I provided resources in the list of references. I also declare that there was no violation of any patents and copyrights during the study and writing of this thesis.

07.02.2022
Hüseyin Okan Soykam

PREFACE AND ACKNOWLEDGEMENT

First and foremost, I would like to thank my supervisor Assoc. Prof. Dr. Günseli Bayram Akçapınar for her mentorship, guidance, support and patience. This work would not be initiated and could have been done without her. Her compassion and tolerance inspire me to become an ethical and qualified academician like her.

I am grateful for my parents; my mother Hanife Nur Tuncer and my father Münir Soykam for their unconditional, endless support to me, my education and my career. I consider myself very lucky to get born to and raised by such family. And, I want to give thanks to my aunt Prof. Dr. Emine Neşe Tuncer for her motherliness and her devoted support to me personally ever since I have come to Istanbul for my education. Lastly, I want to thank my sister Elif Soykam for her unconditional support to me during my study. I would like to thank Şeyma Çokaloğlu Özkaya and Serbulent Ünsal for their friendship and support to the project and to my thesis.

I want to thank Prof. Dr. Osman Uğur Sezerman for his early mentorship and support in my research career before meeting with my current supervisor, Assoc. Prof. Dr. Günseli Bayram Akçapınar. I would also like to thank Assoc. Prof. Tunca Doğan and Prof. Dr. Osman Uğur Sezerman for accepting to participate in my thesis defense as jury members.

Lastly, I would also like to thank to my colleagues and friends in Protein Engineering and Biotechnology Lab (PEBL); Kaan Biçici, Sevde Nur Can, Yunus Emre Dilek, Betül Gündoğdu, İrem Kara, Sena Kıvrak, Hande Osmanağaoğlu, Arda Örçen, Busel Özcan, Şeyma Çolakoğlu Özkaya and Can Veysel Soroğlu for their friendship and support. Although COVID-19 pandemic had emerged before we could physically gather together, I believe, under guidance of our PI Assoc. Prof. Dr. Günseli Bayram Akçapınar, we managed to create a healthy and friendly environment to work together and I am grateful for that. This work is funded by TUSEB Project TA-2019-02-4002.

TABLE OF CONTENTS

LIST OF ABBREVATIONS

ADAM	A Database of AntiMicrobial Peptides
AMP	Antimicrobial Peptide
ANN	Artificial Neural Network
APD	Antimicrobial Peptide Database
AUC	Area Under Curve
CNN	Convolutional Neural Network
CV	Cross Validation
DRAMP	Data Repository of AntiMicrobial Peptides
GRU	Gated Recurrent Unit
LAMP2	Linking of AntiMicrobial Peptides
LSTM	Long-Short Term Memory Network
MCC	Matthew-Correlation Coefficient
RNN	Recurrent Neural Network
SOTA	State of the Art
UniProt	Universal Protein Databank
UniMES	UniProt Metagenomic and Environmental Sequences

LIST OF FIGURES

LIST OF TABLES

ÖZET

Antimikrobiyal Peptit Tanıma için Derin Öğrenme

Patojenik mikroorganizmalar, antibiyotiklere ve diğer antimikrobiyallere karşı hızla direnç geliştirmektedir. Yeni etkili antimikrobiyal terapötiklerin araştırılması, geliştirilmesi, piyasaya sürülmesi zaman almaktadır. Antimikrobiyal peptitler (AMP), patojenik mikroorganizmaları bertaraf etmek için nispeten ucuz, hızlı terapötik alternatiflerdir. Alanda gerçekleştirilen araştırma sayısı çoğaldıkça, son on yılda kamuya açık veri tabanlarında daha fazla veri birikmiştir. Bu durum, peptit tanıma problemini dizi girdisiyle çözmek için makine öğrenimini kullanan hesaplama yöntemlerinin kullanılmasını mümkün kılmaktadır. Son zamanlarda, derin öğrenme yaklaşımlarının tanıma probleminin performansını iyileştirdiği gösterilmiş, ancak bunlar çoğunlukla sınırlı veri setlerin üzerinde gerçekleştirilmiş çalışmalardır. Bu çalışmada, çeşitli antimikrobiyal peptitleri tanıma ve sınıflandırmada yüksek performans gösterebilen bir derin öğrenme modelini eğitmek için daha büyük ve heterojen bir veri seti oluşturuldu. Eğitilmiş model daha sonra Uniprot, *Bacillus subtilis* ve *Trichoderma reseei* proteomlarından türetilen fragman dizilerinin antimikrobiyal aktivitesini tahmin etmek için kullanıldı. Öngörülen AMP adayları, net yük, hidropati, ikincil yapı özellikleri, yapısal kararsızlık indeksi gibi biyokimyasal ve elektrofiziksel özelliklerine göre önceliklendirildi ve sıralandı. Sonuçlarımız, eğitilmiş modelimizin, bir peptit parçasına antimikrobiyal öznitelikleri öğrenerek, eğri altında kalan alan, kesinlik, geri çağırma ve MCC performansı ile makul adayları tahmin edebildiğini göstermektedir. Modelin, geniş bir spektrumda hidropatikliğe, net yüke, ikincil yapıya, motif çeşitliliğine sahip ve antimikrobiyal özellik gösterebilecek AMP adaylarını sınıflandırabildiğini ve bu nedenle tahmin edilen peptitlerin makul AMP adayları olabileceğini gösterdik. Çalışmamız, derin öğrenme modellerinin büyük ve karmaşık bir veriden faydalanarak yüksek boyutlu öznitelikleri öğrenebildiğini ve bu tür derin temsil çalışmalarının AMP tanıma görevi performansını iyileştirme konusunda umut vaat ettiğini göstermektedir.

Anahtar Sözcükler: Derin öğrenme, biyoenformatik, antimikrobiyal peptid, Trichoderma reesei, Bacillus subtilis

ABSTRACT

Deep Learning for Antimicrobial Peptide Recognition

Pathogenic microorganisms are developing resistance to antibiotics and other antimicrobials rapidly. Researching new robust antimicrobial therapeutics to the market is very time consuming and expensive. Antimicrobial peptides (AMP) provide a therapeutic alternative to kill pathogenic microorganisms. As more and more studies are conducted, more data has been accumulating in public databases over the last decade. This enables computational methods to automate the peptide recognition problem via sequence alone. However, sequence-based algorithmic methods have certain limitations and their performance is *subpar*. Recently, deep learning approaches have been suggested to improve the performance over recognition problem yet they were mostly applied to limited datasets. In this study, we curated a bigger and highly heterogenous data set to train a deep learning model that can perform well on a diverse set of AMPs. The trained model is then used to predict the antimicrobial activity of fragment sequences derived from Uniprot, Bacillus subtilis and Trichoderma reseei genomes. Predicted antimicrobial peptide candidates are prioritized and ranked according to their biochemical and electrophysical features such as net charge, hydropathy, secondary structure characteristics, instability index. Our results indicate that our trained model can predict plausible candidates with a reliable AUC, precision, recall and MCC performance by learning the features which makes a peptide fragment to have antimicrobial activity. We have demonstrated that our model can identify AMP candidates with wide range of characteristics such as charge, hydropathy, and with a rich set of BlastP hits pointing to variety of. Our study shows that deep learning models are able to learn useful high dimensional features from a large and complex data and that such deep representation holds promise for improving the AMP recognition task performance.

Keywords: Deep learning, bioinformatics, antimicrobial peptides, Trichoderma reesei, Bacillus subtilis

1 INTRODUCTION AND AIM

Infectious diseases have always been a threat to humanity throughout history. These diseases even had profound impact on human migration patterns so that particular populations tried to evade pandemics. Despite human effort to evade and run away from pandemics, there are many infamous yet disastrous pandemics which affected and killed humans at massive scale (1,2). For example, a prehistoric pandemic about 5.000 years ago believed to be responsible for wiping out an entire prehistoric village in China (3). Plague of Justinian has ravaged Byzantian Empire (4) and historians regards it as one of the primary factors that started the decline of Byzantian Empire (5). The infamous Black Death traveled from Asia to Europe and it is estimated to have wiped out over half of European population. It even contributed to the change within the course of Europe's history that labor became harder to find and serfdom ended (6,7).

Fast forward to modern industrial age, flu pandemic started in St. Petersburg and rapidly spread across Europe, killing 1 million people despite its 1-year time span (8). It is known that it took only five weeks to reach peak mortality (9). Similarly, Spanish Flu was estimated to take 500 million lifes sweeping from South Sea to North Pole (10). Chronologically after Spanish Flu, the first antibiotic called penicillin has been discovered by Alexander Fleming in 1928 (11). Although, it was a revolutionary step towards fighting against microbial pathogens, these particular antibiotics has started to become swiftly obsolete as microbial organisms against which these antibiotics are administered are becoming resistant to them through evolution (12,13). This implies investigating and developing new potential effective antibiotics must be an ongoing endeavor in order to replace old obsolete antibiotics with new effective ones.

Since researching, developing and deploying new robust antibiotics is very time consuming and expensive (14), this makes them ineffective as the latency for deploying new therapeutics to the market is not compatible with the speed at which microorganisms developing resistance to them (15). In addition to antibiotics, antimicrobial peptides provide an effective alternative therapeutic potential to kill pathogenic microorganisms (16,17).

Developments in statistics, computer science and hardware technology sparkled many advancements in machine learning which is a sub field of artificial intelligence study. These developments enabled the development of many successful methods for superior performance. Particularly, with recent advancements in Graphical Processing Units (GPUs), a new branch has emerged (namely reborned) from machine learning named deep learning. Deep learning provided the methodological and systematic utility for researchers to leverage advancements in hardware technology. Many recently developed deep learning architectures and methods proved to be having best performance on many tasks ranging from computer vision (CV) to natural language processing (NLP) and bioinformatics. However, one practical disadvantage of deep learning is that it requires high amount of quality data compared to shallow (traditional) machine learning algorithms. Since human labelling is required for almost all cases in order to generate quality data, it requires significant amount of time to produce high quality data in high quantity which is necessary to train deep learning models. This makes deep learning harder to apply on some domains medicine alike. Fortunately, as more and more studies are conducted, more data has been accumulating in public antimicrobial databases over the last decade. In return, this opens up an opportunity to apply deep learning techniques to antimicrobial peptide recognition problem which holds promise to surpass the performance of shallow machine learning models established in the literature.

In this study, we test the hypothesis that novel deep learning methods and/or architectures can improve the performance on classifying antimicrobial peptides. That is, we explored how several state-of-the-art deep learning methods performs, compared their performance and attempted to provide plausible explanation on the performance of these particular architectures. To this extent, we aim to develop an end-to-end method for recognizing antimicrobial peptides by using sequence information solely. Moreover, we have several goals which we think can improve current state of literature and benefits researchers. Our first goal is to curate a more heterogenous and comprehensive antimicrobial peptide dataset which could possibly reflect the real life situation. Up to now, most of the research on the application of machine learning algorithms (and deep learning algorithms) and their successful applications heavily rely on the use of datasets formed from relatively homogeneous subclasses of antimicrobial peptides. Secondly, we have a goal of fair benchmark of current state of

the art deep learning methods on this dataset. Finally, our last goal is acquiring insights of how and in which manner sequence information provides utility in classifying antimicrobial peptides from statistical and methodological perspective.

2 BACKGROUND

2.1 Antimicrobial Peptides

Antimicrobial peptides (AMPs) are the host defense peptides which plays role in immunity and host defense. AMPs are most likely to be cationic (positively charged) and amphiphilic (hydrophilic and hydrophobic) α-helical molecules.

Typically, AMPs contain between 10 to 60 amino acid residues with their average length being 33.26, and they have an average net charge of 3.32 (18). Despite their positive average net charge, several anionic AMPs also exists. And they contain some acidic amino acids like aspartic acid and glutamic acid (19–21).

AMPs can be classified according to several aspects such as their biological source, biological activity, amino acid richness-like characteristics and structure. Another important criterion for classification is whether AMPs form disulfide bridges (22). Disulfide-bridged AMPs (DBAMPs) usually targets the membrane boundaries of ion channels. They were used in treatment of several autoimmune and neurological diseases such as epilepsy, glioma etc. (23,24) and they can be further classified into four subgroups based on their mechanism of ion channel targeting. However, this classification scheme along with DBAMPs are not a particular focus of this study as such they won't be detailed. It is important to note that these aforementioned classification ontologies are not mutually exclusive, they coincide within each other as they represent distinct point of views on same AMPs.

2.1.1 Classification of antimicrobial peptides based on sources

The sources of AMPs can be divided into mammals, amphibians, microorganisms, and insects according to the accumulated data in APD3 (25). Human host defense peptides constitute the largest portion of mammalian AMPs, meanwhile there has been an increased attention for AMPs found in oceans, from ocean species however data accumulated in this area is not ample and therefore they are not classified and described as a distinct source in this thesis.

2.1.1.1 Mammalian antimicrobial peptides

Mammalian AMPs are the antimicrobial peptides found in human, sheep, cattle, and other vertebrates. Cathelicidins and defensins are the two main families of the mammalian AMPs and also can be classified under host defense peptides. Cathelicidins are characterized by their cathelin-like domain meanwhile defensins are characterized by the disulfide bonds present in between their molecules.

Cathelicidins are small, cationic antimicrobial peptides with amphipathic properties and they take role in the main component of the innate immune system in many vertebrates, including humans (26). Their role is part of the oxygen-independent activity against pathogens (27). A lot of cathelicidins, similar to lot of mammalian antimicrobial peptides, are stored in neutrophils and granules. The peptide family named cathelicidins with a common proregion (cathelin domain) were first identified in mammals in bone marrow myeloid cells.

The defensins represents a class of cationic antimicrobial peptides that play a vital role in innate and adaptive immunity. Defensins are a major family of host defense peptides expressed predominantly in neutrophils and epithelial cells. They partake in defense mechanisms against viral infection and bacterial infection as well as in tumorigenesis since most cancers develop from epithelial cells and tissues. Defensins can be divided into α-, β-, and θ-defensins based on their disulfide bridge conformations.

2.1.1.2 Amphibian-derived antimicrobial peptides

Antimicrobial peptides produced by amphibian species constitutes another major source of AMPs had been discovered so far. Amphibian AMPs plays a crucial role in protecting the amphibian species from pathogenic organisms which they encounter in their natural environment. Out of the amphibian species, frogs are the primary source of amphibian AMPs considering the quantity and diversity of amphibian AMPs they can produce. One of the most famous frog AMPs is magainin which is found in *Xenopus laevis* (African claw frog) and produced by mucous glands in the skin of *X.*

laevis. Magainin shows a broad-spectrum of antimicrobial activity against both gram-positive and gram-negative bacteria along with fungi and protozoa. Magainin has also been shown to demonstrate anticancer properties in a setting where they had been tested against human melanoma cell lines. By this, they are the first AMPs produced by a multi cellular organism to be pivoted for therapeutic uses.

Temporins are another species of AMPs which are produced by granular glands of frogs belonging to the genus of *Rana.* They are generally secreted upon an incident of injury to protect the wound against external infections of pathogens. Not only they exert antimicrobial activity against particular species of Gram-positive bacteria, when applied in higher concentration levels, they can demonstrate antiparasitic effects against certain parasites such as *Leishmania* too. Temporins are typically 8-17 AA residues long.

2.1.1.3 Insect-derived antimicrobial peptides

Due to their strong adaptive nature which enables insects to survive, insects are another major source of AMPs in nature. Insects mainly synthesize their AMPs in their fat body tissues and blood cells. Cecropin family of AMPs, being the most famous family of insect AMPS, are first isolated from the hemolymph of *Hyalophora cecropia*. Cecropin family of peptides shows antimicrobial activity against Gram-positive and Gram-negative bacteria in addition to their antifungal properties. Despite insects produce many AMPs for their survival purposes, the number of AMPs discovered so far varies greatly between species. Some species such as *Harmonia axyridis* (invasive harlequing ladybird) and *Hermetica illucens* (black soldier fly) produces up to 50 different AMPs whereas *Acyrthosiphon pisum* (pea aphid) lacks AMPs or some peptides like Jellein are recently being discovered from bee royal jelly.

2.1.1.4 Microorganism-derived antimicrobial peptides

Through the course of evolution, as much as multi cellular organisms developed protection mechanisms against pathogens, microorganisms too developed mechanisms which provide them advantages for the competition between them and the pathogens.

Therefore, it is found that many bacteria and fungi synthesize AMPs to eliminate their competitor microorganisms in their environment. Some widely known examples are nisin, gramicidin from *Lactococcus lactis, Bacillus subtilis, and Bacillus brevis.* Since the chemical synthesis of similar peptides are costly, researchers have turned to obtain naturally produced peptides from microorganisms. For this reason, particular fungal species such as *Pichia pastoris, Trichoderma reseei and Saccharomyces cerevisiae* and bacterial species like *Escherichia coli* and *Bacillus subtilis* have been used as recombinant expression hosts. It is important to note that because of certain important aspects such as host toxicity, proteolytic degradation and purification, expression of AMPs can be challenging and daunting in these recombinant expression systems.

2.1.2 Classification of antimicrobial peptides based on activity

Although there are many classification systems defined by various researchers and databases, APD3 divides AMPs into 18 categories. These fine-grained categories can be further abstracted up and summarized into much broader ones as antibacterial, antifungal, antiviral, antiparasitic and anticancer (anti-tumor) peptides according to which kingdom of organisms they are effective against.

2.1.2.1 Antibacterial peptides

A significantly large portion of AMPs are antibacterial AMPs which can be derived from many sources. These antibacterial AMPs have either an inhibitory or a lethal effect for pathogenic bacteria.

2.2 Antimicrobial Peptide Recognition Problem

Antimicrobial peptide recognition problem is a classification problem in which an input sequence X from an arbitrary fragment is determined to be whether having antimicrobial activity or not. In this study, the general recognition problem is reduced to a binary classification problem where a hypothesized model classifies a given peptide sequence as either AMP or decoy (not AMP). The problem is tried to be solved by training supervised models using well known AMP and decoy peptides in the literature. That is, the weights of any hypothesized model parameters are optimized such that the best deep representation is achieved by reducing the performance error using several performance metrics.

2.3 Deep Learning

Machine learning is a field of vast artificial intelligence (AI) research which aims to create sentient and cognitively advanced forms of intelligence system able to mimic advance human behavior. The beginning of AI research has been established by classical philosophers to describe and formalize human cognition as a symbolic system until it was formally founded at a conference at Dartmouth College, in Hanover, in 1956 by coining the term “artificial intelligence”.

Since first coining of the term AI, there have been numerous approaches to create AI such as probabilistic programming, expert systems, formal logic and statistical learning. As the data generated in the world became abundant and ubiquitous, researches started explore machine learning approaches which tries to recognize patterns in data via curve fitting by statistical methods. One particular approach of machine learning is deep learning which is based on artificial neural networks created to mimic biological neural network found in brain. These artificial neural networks despite first created and modelled as a 3-layer shallow networks, can also be constructed with many numbers of layers which is how the term deep used to distinguish deep learning from classical (shallow) machine learning.

Although there has been interest in artificial neural networks and deep neural networks for many years, it had never seen the progress it saw today due to several factors. These factors are; i) amount and quality of labelled data ii) computing power suitable to train deep neural networks and iii) efficient training algorithms particularly tailored to train those networks. Breaking off the predicted threshold depicted by Moore’s Law and recently emerged GPU technologies along with the massive amount of surge in labelled data motivated and enabled the researchers to return to deep neural network research again. With the invention of efficient gradient boosting algorithms to train deep neural networks on GPUs in a parallel fashion, these deep networks, starting with computer vision applications, have demonstrated much promise with their glaring performance on fundamental computer vision task such as image classification, object recognition etc.

Despite seemingly overfit in many validation scheme, deep learning models tend to generalize better and perform better than most classical machine learning algorithms. Researchers have found that deep learning models trained using gradient boosting with sufficient data can escape local maximas and saddle points more easily from an optimization perspective. And that, they can learn and represent the manifold of the search space in which data and the related task contained within. These enable deep learning to perform and generalize better than most alternative approaches given there exists sufficient amount of high quality labelled data for the related task.

2.3.1 Artificial neural networks

The most basic form of Artificial Neural Networks (ANN) are feed forward neural network in which there is one input layer, one intermediate layer and one output layer. The input layer takes in the input corresponds to fed data in numerical form. The output layer relates to the output numerical value of predicted or ground truth label for any given data point. The middle layer is then computed via gradient boosting of errors between predicted and ground truth output values for all training data points such that the weights of the middle layer will be optimized for best overall performance for the targeted performance metric. This enables the ANN model to arrive at a reduced representation of the search space as to best provided performance for the given task.

2.3.2 Convolutional neural networks

Convolutional neural networks (CNN) introduces a two novel distinct layer types and corresponding operations to the basic ANNs. The first operation convolution is first introduced in computer vision field for the image classification task. The convolutional layer scans the input sequence in a sliding window fashion called filters with fixed size with a fixed stride length and each filter computes a convoluted metric for all the input amino acids in filtered window as a single metric standing for a composed unit. The accumulated convoluted metrics then fed to the pooling layer to calculate a summary metric out of all the convoluted metrics from the nodes of the convolution layer to calculate a summary metric which represent a high level compositional feature vector of the shared parameters of the previous convolutional layers. Pooled vectors are then linearized to fed to the fully connected feed forward

layer to calculate the interaction between those high level features. The weights of the feed forward layer are then feed to the final activation layer to calculate the final output label for the given sequence.

2.3.3 Bidirectional long-short term memory networks

Long-Short Term Memory Networks (LSTM) are introduced model the time variant information hidden in the input space as an improvement to recurrent neural networks.

The memory module in LSTMs holds a cell state information in order to alleviate the long-term dependency problem encountered by RNNs. That is, RNNs are not as successful as LSTMs in modelling the interactions of features that are distant to each other in the sequence space. LSTMs are able to this by using the gate mechanism to determine whether it remembers or forget a modelled interaction using the provided feedback by the descended error. Gates are used to optionally let information let through the layers to do so.

2.3.4 Stacked long-short term memory networks

LSTMS are stacked to further and further long-term dependency modelling of the input in order to improve performance.

2.3.5 Convolutional neural network-recurrent neural networks

Instead of directly modelling the input sequences, the convoluted feature representation of the sequence is fed to the RNNs and the time-variant interactions are modelled in a high dimensional vector space.

2.3.6 Convolutional neural network -gated recurrent units

Instead of directly modelling the input sequences, the convoluted feature representation of the sequence are fed to the GRUs with a gate mechanism and the time-variant interactions are modelled in a high dimensional vector space.

2.3.7 Convolutional neural network-long-short term memory networks (Veltri et. al.)

Instead of directly modelling the input sequences, the convoluted feature representation of the sequence are fed to the LSTMs with memory mechanism and the time-variant interactions are modelled in a high dimensional vector space.

2.3.8 Convolutional neural network-bidirectional long-short term memory networks

Instead of directly modelling the input sequences, the convoluted feature representation of the sequence are fed to the LSTMs with memory mechanism forward and backward fashion, and the time-variant interactions are modelled in a high dimensional vector space.

3 MATERIALS AND METHODS

3.1 Databases

3.1.1 ADAM (a database of antimicrobial peptides)

ADAM is a large-scale comprehensive antimicrobial database, containing 7.007 unique sequences and 759 structures. It aims to provide a comprehensive source on established associations between AMP sequences and their structures via linking them to structural folds. It contains 30 distinct AMP structural fold clusters that has more than one structural fold and around 1000 AMPs linked to an at least 1 structural fold. The cumulative of these associations implies that AMPs only cover 3% of the overall protein fold space as defined by SCOP and CATH according to the analyses performed within ADAM.

7007 unique sequences are collected from twelve established databases and assigned a unique ADAM ID. These are APD2, AVPpred, BACTIBASE, BAGEL3, CAMP2, DAPD, DAMPD, HIPdb, PenBase, PhytAMP, RAPD, and YADAMP. The AMP sequences contained in ADAM were mostly derived from natural sources, from organisms such as archaea, bacteria, plants and animals while 2497 out of the 7007 sequences have been experimentally validated as per recorded in the literature.

The AMP structures contained in ADAM were obtained by running BLAST for the experimentally validated AMP sequences against the Protein Data Bank. 408 sequences had 759 matching structures with either 100% sequence identity with any E value or greater than 90% sequence identity with E-value lesser than 10^-5. Next, each matching structure was annotated by SCOP v1.75B and CATH 4.20 simultaneously. Since some AMP structures did not have any CATH or SCOP annotations, all unique AMP structural could not been determined using only CATH and SCOP annotations.

After obtaining the sequences and structures, a graph-based clustering algorithm was used for establishing unique AMP folds. In the graph configuration, the vertices represent AMP structures and the edges in between them represent similarity in

between these structures. This similarity is measured by TM-score and an edge exists if the TM score between two structures are greater than 0.5 which indicates that these two structures should belong to the same fold. 136 AMP fold clusters were formed with 30 clusters containing more than one AMP structure where each AMP is allowed to have only 1 best matching structure, and multiple AMPS can map to the same AMP structure.

AMP sequence to structure is established by performing BLAST on the experimentally validated AMPS against PDB. Over 1000 AMP sequences were found to matched against an AMP structure which is determined as the best maching AMP structure with E-value being lesser than 10^-5.

The database and the accompanying website were built using AppServ 2.6.0 using The Apache HTTP server as the backend server, MySQL as the database engine, and PHP is used for server-side scripting. The website and the database can be accessed through http://bioinformatics.cs.ntou.edu.tw/ADAM/index.html.

3.1.2 APD3 (the antimicrobial peptide database)

APD is an antimicrobial database focusing on AMPs from natural sources with known sequence and proved activity. It initially went online in 2003, updated in 2009 for version, and the current version APD3 is further expanded in 2013.

APD3 contains 3257 AMPs from six kingdoms including 365 bacteriocins/peptide antibiotics from bacteria, 5 from archaea, 8 from protists, 22 from fungi, 360 from plants and 2414 from animals, plus some additional synthetic AMPs with several activities. These activities which also corresponds to subtypes of AMPs are antibacterial peptides, antibiofilm peptides, anti-MRSA peptides, anti-TB peptides, anti-endoxin peptides, anti-toxin peptides, antiviral peptides, anti-HIV peptides, antifungal peptides, anti-candida peptides, antiparasitic peptides, anti-diabetic peptides, wound healing peptides, chemotactic peptides, anti-inflammatory peptides, spermicidal peptides, insecticidal peptides, ion channel inhibitors, protease inhibitors, antioxidant peptides, surface immobilized peptides, two-chain peptides, and synergistic peptides.

The peptide sequences for the initial version (published in 2004) were collected from literature by searching PubMed using keywords such as ‘antimicrobial peptide’, ‘antibacterial peptide’, ‘antifungal peptide’, ‘anticancer peptide’ or ‘antitumor peptide’. This resulted in total of 525 peptides. Initially collected peptides were mainly from natural sources.

In 2009, it is updated to APD2 by doubling the amount of its peptide entries, making it 1228 peptide sequences. Statistical analyses performed using the peptide sequences in APD2 database revealed that the frequently used amino acid residues which defined as greater than 10% fraction, are Ala and Gly in bacterial peptides, Cys and Glu in plant peptides, Ala, Gly and Lys in insect peptides, and Leu, Ala, Gly and Lys in amphibian peptides. Using such frequently occurring residues as a basis, APD2 demonstrated database-aided peptide design approaches and GLK-19 of the three designed peptides showed a higher activity against Escherichia coli than human LL-37. Some new features for search interface such searching by peptide name or functionalities to explore peptide families has been implemented to the web server in order to improve user experience.

APD3 expanded upon APD2 by increasing the total amount of peptide sequences to 2619 AMPs with richer defined (with more subcategories) activities in 2016. In addition to the sequence information, each peptide sequence is integrated with additional informations such as chemical modifications, structural classifications, peptide paramaters such as length, charge, hydropathy etc., associated source organism and its taxonomic classification along with additional functions belong to peptides. Search usability is improved by adding a capability to search a peptide using its sequence as it is the easiest way to find a peptide in any data collection.

APD3 are installed on a web server and offers a web interface which is available at http://aps.unmc.edu/AP/.

3.1.3 DBAASP (database of antimicrobial activity and structure of peptides)

DBAASP is a database that has been first established in 2014 with the aim of providing necessary information and analytic resources to the scientific community to develop antimicrobial compounds with high therapeutic index. It contains experimentally validated antimicrobial peptide sequences and offers compete information on chemical structure, target species, target object of cell, and antimicrobial, hemolytic and cytotoxic activities of these peptides coupled with their experimental conditions at which aforementioned activities are tested for.

As in the case of APD, AMPs were collected using search keywords ‘antimicrobial’, ‘antibacterial’, ‘antifungal’, ‘antiviral’, ‘antitumour’, ‘anticancer’ and ‘antiparasitic’ peptides on PubMed. In the same manner, all side information regarding peptides are collected from the publications resulted from the PubMed search. DBAASP contains in total 4054 peptides of which 3917 are monomers, 27 are dimers, 110 are two-peptides.

DBAASP uses IMGT-ONTOLOGY, the unique ontology tailored for immunogenetics and immunoinformatics and borrows its concepts of identification, classification and description to characterize peptides. Identification holds information with respect to name of peptide, synthesis type (ribosomal, nonribosomal, synthetic), completixity (monomer, dimer, two-peptide), target group (Gram-positive, Gram-negative, fungus, virus, cancer etc.), and target object of cell (memberane protein cytoplasmic protein, lipid bilayer and DNA/RNA). Classification involves information with respect to the kingdom source of species as to Cavalier-Smith’s six kingdom system and source species. Description includes sequence, length of peptide and post-translational modifications.

DBAASP is updated into version 2 in 2016 which introduced a unique edge compared to competitor databases such as reported MIC values and chemical structures. It reports chemical structures against more than 4200 specific target microbes for more than 2000 ribosomal, 80 non-ribosomal and 5700 synthetic peptides. 300 unique distinct MD simulations for novel structures are introduced in

this version. MD simulations are performed using the Phyre2 web server. A ranking search model has been added to enable users to rank peptides based on their hemolytic and cytotoxic activity against target cells with the corresponding measure of biological activity.

Latest update, DBAASP v3 was published on 2021. This iteration increases the number of entries and provides assisting tools for developing and optimizing de novo peptides with desired biological functions such as a sophisticated multifactor analysis of selected physicochemical properties. Other tools include PGA tool to predict the general activity of a given peptide, and PAASS tool to predict the antimicrobial activity of a given peptide against particular target strains. Furthermore, DBAASP has implemented a streamline MD simulation upload pipeline to facilitate the process of uploading MD simulations to the database.

DBAASP v3 was built on Spring Boot2.x and runs on a Docker-based Centos 8.x with Java8 Runtime and uses MariaDB as the backend database engine. DBAASP web site can be found at https://dbaasp.org/.

3.1.4 DRAMP (data repository of antimicrobial peptides)

DRAMP, first published in 2016, is an open-access and manually created database harboring diverse annotations of AMPs including sequences, structures, activities, physicochemical, and patent, clinical and reference information. The clinical dataset describes detailed information such as the clinical trial phase, therapeutic applications, company and preclinical or clinical AMP literature sources.

Antimicrobial peptides in DRAMP were collected from PubMed, Swiss-prot and Lens by using keywords such as ‘antimicrobial peptide’, ‘antibacterial peptide’, ‘antifungal peptide’, ‘antiviral peptide’, ‘antitumor peptide’, ‘anticancer peptide’, ‘antiparasitic peptides’ or ‘insecticidal peptide’. Resulting hits are only registered into the database should they conform to the preset criteria: i) their antimicrobial activities have been experimentally validated; ii) the amino acid sequences of peptides have been completely elucidated to contain no ambiguous residues; iii) precursor and signal

regions have been removed to remain only actual activity showing region of peptides; iv) their length is less than 100 amino acid residues.

DRAMP currently contains 22259 entries, 5891 of which are general AMPs, 16110 patent AMPs and 77 AMPs are in drug development phase (preclinical or clinical stage). These two distinct sourced AMPs are divided into two dataset as general and patent datasets. Clinical information was extracted from literature into a separate clinical dataset. Sequence, length, name, relevant database identifiers, MIC values, biological activity field, gene, source, family, binding target, structure type, description, PDB ID, PubMed ID and physiochemical information such as Boman index, mass and half-life annotations can be found for all peptides in the general dataset. In the patent dataset, only sequence, name, source and activity are reported with regarding patent information.

DRAMP v2 published in 2019, added new entries into the database with additional annotations. These are hemolytic activity and the corresponding PubMed ID that provides the demonstration for the biological activity of a peptide. 44 new X-ray crystallography and NMR spectroscopy structures from PDB are appended into the database.

DRAMP was built with Apache web server v2.2.22 with MySEQ v5.5.29 as backend and runs on 32-bit Linux platform. It can be accessed at http://dramp.cpu-bioinfor.org/

3.1.5 LAMP2 (a database for linking antimicrobial peptides)

LAMP aims to provide an integrated open-access database with accompanying tools for assisting antimicrobial agent discovery and design in response to recently ascending problem of drug resistance in pathogens. It links AMPs with their related cytotoxicity along with antimicrobial activity information to package and present them in an integrated manner. More importantly, it links these AMPs and the containing databases together using cross-links to enhance the experience of researcher.

LAMP includes 3.904 natural AMPs and 1.643 synthetic AMPs with a total of 5.547 AMP entries. All the AMPs in LAMP are shorter than 100 amino acids sequences.

3.1.6 Uniprot (the universal protein database and swiss-prot)

UniProt (www.uniprot.org) is the largest and most comprehensive open access protein sequence database with high quality web interface and service. It is established to provide a high quality and free resource on protein sequence and related functional annotations to the scientific community.

3.2 Dataset Curation

AutoAMP dataset was created by collecting AMP and decoy sequences from various public databases and datasets as shown in Figure 1 to create a heterogenous, larger and thereby more challenging dataset. In order to expand the dataset, we added decoy sequences from UniProt database5. We used cd-hit6 to perform redundancy reduction with 90% similarity to achieve final dataset (10591 AMP and 23690 DECOY sequences).

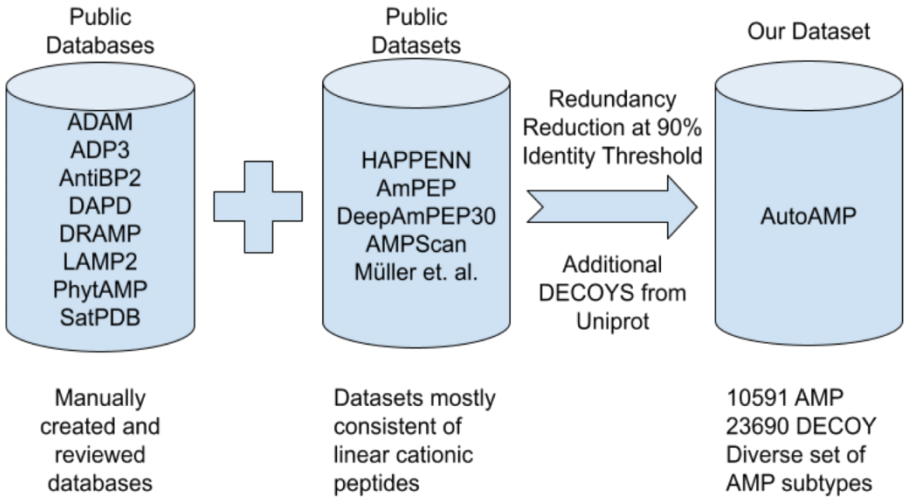


Figure 1. Overview of dataset curation.

3.3 Deep Learning

3.3.1 Convolutional neural networks

An embedding layer of 128 dimension is used to project the input to a higher dimensional space. The embedding layer is followed by two consequent convolutional layers with 64 filters, kernel size of 16 and 1 stride size. Dropout of 0.5 dropout probability is used before the 1D max pooling layer with pool size of 5. Finally, pools are flattened and connected to the final feed forward output layer with sigmoid activation function.

3.3.2 Bidirectional long-short term memory networks

An embedding layer of 128 dimension is used to project the input to a higher dimensional space. The embedding layer is followed by Bidirectional LSTM sequential modelling layer with 100 nodes and 0.1 probability of dropout with unrolling enabled. BiLSTM layer is then connected to the final feed forward output layer with sigmoid activation function.

3.3.3 Stacked bidirectional long-short term memory networks

An embedding layer of 128 dimension is used to project the input to a higher dimensional space. The embedding layer is followed by 3 stacked Bidirectional LSTM sequential modelling layers with 100 nodes and 0.1 probability of dropout with unrolling enabled. Final, BiLSTM layer is then connected to the final feed forward output layer with sigmoid activation function.

3.3.4 Convolutional neural network-recurrent neural networks

An embedding layer of 128 dimension is used to project the input to a higher dimensional space. The embedding layer is followed by two consequent convolutional layers with 64 filters, kernel size of 16 and 1 stride size. Dropout of 0.5 dropout probability is used before the max pooling layer with pool size of 5. Max pooling layer

then flattens and connects to the RNN modelling layer before finally connected to the final feed forward output layer with sigmoid activation function.

3.3.5 Convolutional neural network-gated recurrent units

An embedding layer of 128 dimension is used to project the input to a higher dimensional space. The embedding layer is followed by two consequent convolutional layers with 64 filters, kernel size of 16 and 1 stride size. Dropout of 0.5 dropout probability is used before the max pooling layer with pool size of 5. Max pooling layer then flattens and connects to the GRU modelling layer before finally connected to the final feed forward output layer with sigmoid activation function.

3.3.6 Convolutional neural networks-long-short term memory networks (Veltri et. al.)

An embedding layer of 128 dimension is used to project the input to a higher dimensional space. The embedding layer is followed by a convolutional later and a max pooling layer. Max pooling layer then flattens and connects to the LSTM modelling layer before finally connected to the final feed forward output layer with sigmoid activation function. The graphical illustration of the architecture can be found in Figure 2.

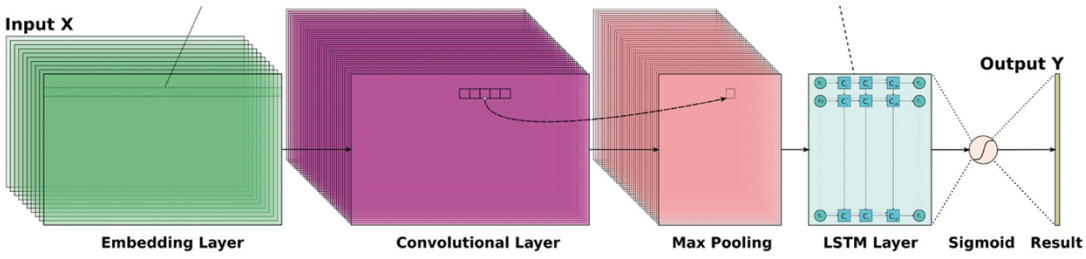


Figure 2. Architecture of Veltri et. al. CNN-LSTM.

3.3.7 Convolutional neural network-bidirectional long-short term memory networks

An embedding layer of 128 dimension is used to project the input to a higher dimensional space. The embedding layer is followed by two consequent convolutional layers with 64 filters, kernel size of 16 and 1 stride size. Dropout of 0.5 dropout probability is used before the 1D max pooling layer with pool size of 5. Finally, pools are flattened and connected to the final feed forward output layer with sigmoid activation function.

3.4 Model Training and Validation

Curated AutoAMP dataset is randomly split into two partitions: training+validation and a heldout. Training+validation set constitutes of %80 of the whole dataset whereas %20 of the dataset is used for held out test set. Training+validation set is then used to train and validate several models in 10-fold cross validation scheme. That is, the partition is splitted into 10 shuffled partitions and 1 partition is randomly selected for validation and the remaining 9 partitions are used for training to estimate performance of a given model. This whole process is then repeated 10 times as in 10 folds, and the resulting performance metrics are estimated by calculating the means and standard deviations of the metrics across all folds. The shuffling of partitions in every fold is performed in a stratified manner to ensure the class balance of every fold reflects the class balance of each dataset. Doing such 10-fold stratified cross-validation enables the estimation of the model performance with less bias and more variance so long as the dataset is large enough which is true in this case.

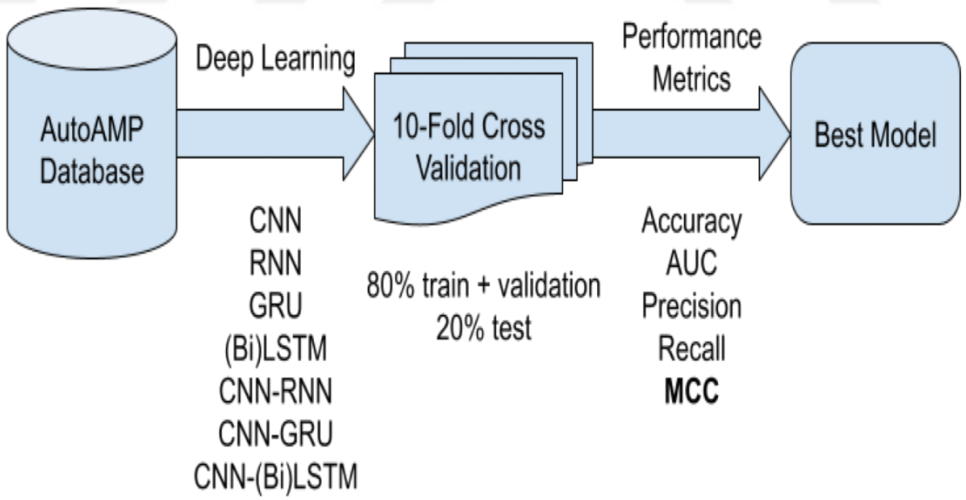


Figure 3. Model training and validation flowchart.

3.5 Peptide Prediction and Prioritization

The model with the best performance suitable for our practical goals were selected and then deployed to scan and predict the relevant class of a given peptide sequence either as AMP or a decoy. We hypothesize that from an information guided perspective, there can be many peptide-length sequences which can potentially behave as AMPs due to their sequential or structural make up, and that from a biological perspective, some peptides related to histones, transcription binding factors, immune system related peptides etc. can be found in either as part of proteins or in the CDS sequence of the relevant protein. For this reason, we developed a two-step algorithm to generate potential N-length fragments from a given sequence in sliding windows over the given input sequences to generate potential peptide fragments. Generated fragments are first subjected to redundancy reduction filtering and then the remaining fragments are predicted by the model as AMP or decoy. The candidates with a prediction score of 1 are then used as inputs to our prioritization procedure and the top 10 ranked peptides are inspected for further analyses. An exemplary overview of the generate-predict-prioritize process can be seen in Figure 4.

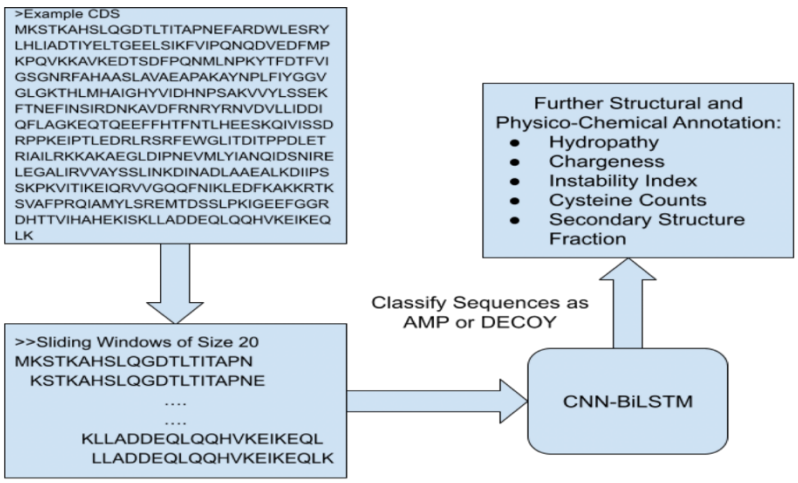


Figure 4. Overview of generate-predict-prioritize process applied to *Trichoderma reseei* CDS sequences as an example.

Biochemical and electrophysical properties of peptides are used as features to filter and rank the peptides within each other. Prioritization was necessary because even the number of peptides with a prediction score of 1 were in impractical amounts for manual inspection. For this reason, the steps and principles outlined below are used in the decision-making process of filtering and ranking the peptides:

1. It is known that AMPs are bias towards being positively charged in general. For this reason, peptides with relatively higher net charges should be prioritized.
2. Hydropathicity plays a crucial role in AMPs' anti-microbial activities. As such, peptides with relatively higher hydropathy scores should be prioritized.
3. Peptides with both higher net charges and hydropathy scores should be ranked on top of the prioritized peptides for manual inspection.
4. Cystein motifs can be seen in many AMPs, especially in those of antifungals. Even though, it is harder to synthesize cysteine rich peptides, they are prioritized and ranked within each other as a separate set.
5. Some candidate peptides, despite having desirable biochemical and electro-physical features for an AMP, predicted to have higher protein instability. With practical concerns considered, these aforementioned candidate peptides with an instability score of higher than 40 are filtered out and had not been considered for further analysis.

Generate-predict-prioritize workflow was repeated for UniProt data dump, *Triocherma reseei* CDS and *Bacillus subtilis* CDS such that the window length parameter is adjusted based on the contextual biological information. In order to determine the optimum window lengths, mean values from exploratory analysis is used for UniProt data dump, and mean values of AMPs produced by *T. reseei* and *B. subtilis* to scan their corresponding CDS sequences.

4 RESULTS

4.1 Dataset Analysis

The properties of the curated dataset such as amino acid frequency, length distributions for both AMP and Decoy sets are shown in Figure 5 and Figure 6, respectively. Amino acid fraction densities in Figure 5 suggest that there is not a distinct motif which can discern between AMPs and decoy peptides. AMPs have more Cysteine, Leucine, Glycine, Tryptophan and Lysine whereas decoy peptides have more Valine, Glutamic Acid, Aspartic Acid, Methionine and Threonine. The magnitude of density differences of aforementioned AAs vary in between and rest of AAs are more or less equally represented in both AMPs and decoy peptides. This suggest that sole existence of AAs in a given sequence is not enough to be used to determine the antimicrobial activity.

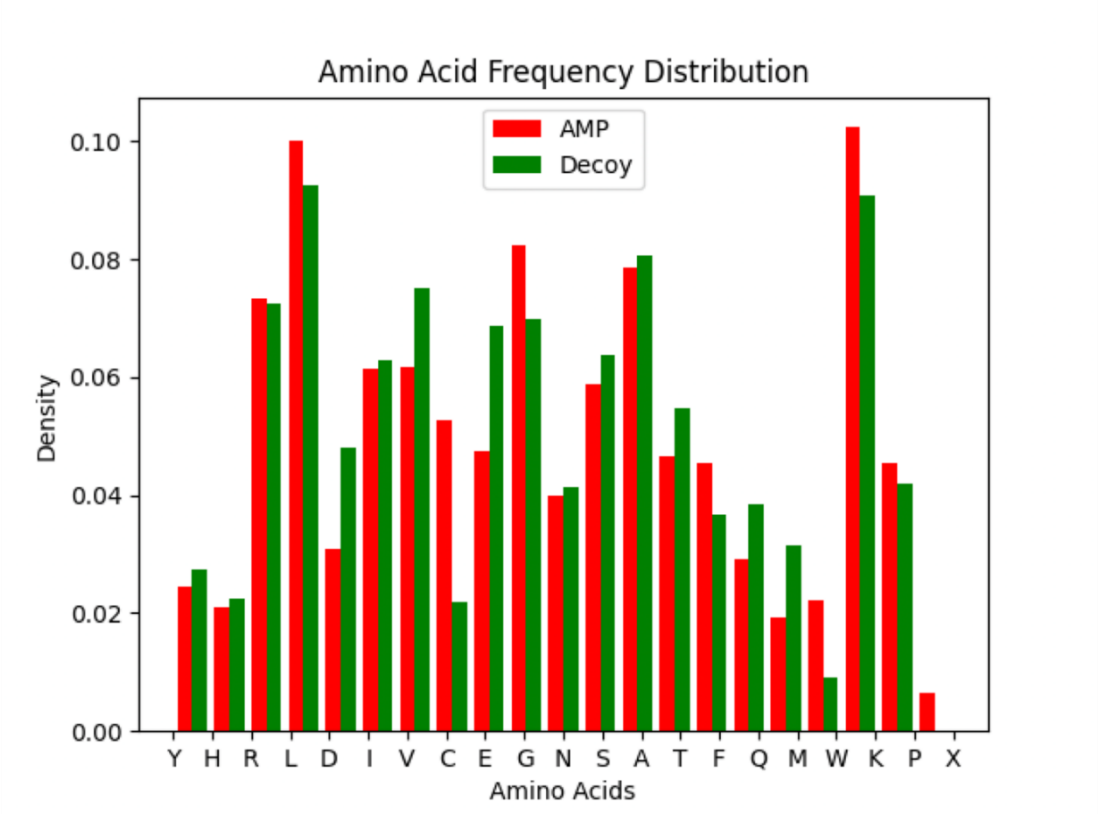


Figure 5. Amino acid frequency distributions of AMPs and decoy peptides.

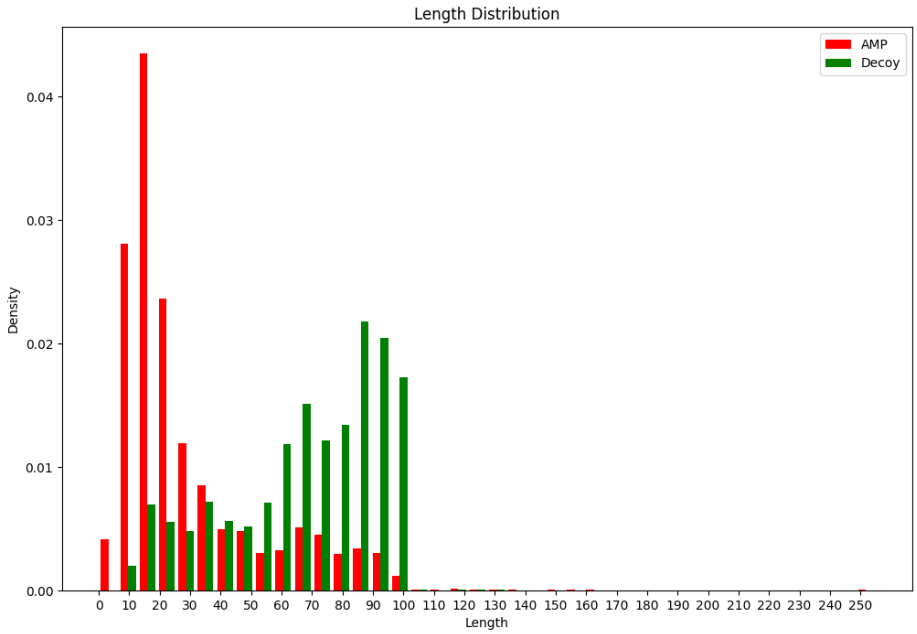


Figure 6. Length distribution of AMPs and decoy peptides

Next, we analyzed the length distributions of AMPs and decoy peptides given in Figure 6 in order to determine if there is a tendency towards certain peptide lengths in nature. Peptides lengths between 10 and 30 covers significant number of peptides while decoy peptides are much longer compared to AMPs, and that, AMPs that are smaller than 10 AAs constitute less than %1 percent of all AMPs in the dataset. Based on these exploratory findings, we filtered out all AMPs which are shorter than 10 and we used various window lengths of 15, 20 and 30 for scanning and prediction in order to find more peptides in limited amount of time.

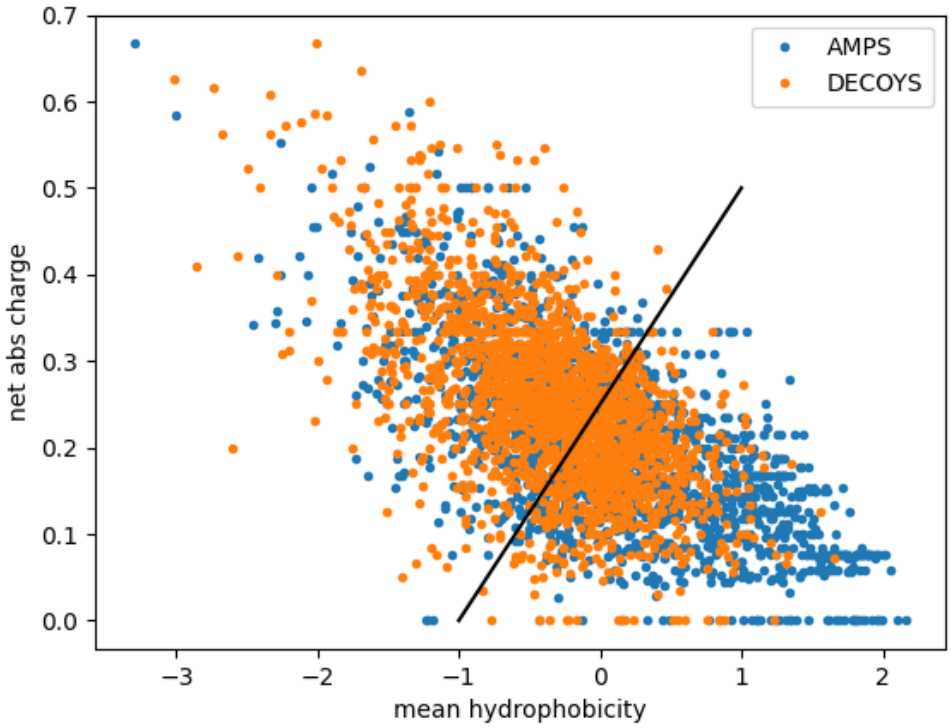


Figure 7. Uversky plot of AMPs and decoy peptides.

A Uversky plot, exhibiting distribution of the hydrophobicity and net absolute charge properties of the peptides in both AMP and decoy peptide datasets is shown in Figure 7. As it can be seen in the Uversky plot, AMPs and decoys cannot be discerned solely using simple features such as net absolute charge or mean hydrophobicity. This suggest that more complex, high level patterns of evolutionary and structural information is hidden in the sequence which can be exploited by multilayered deep learning architectures.

4.2 Performance of Deep Learning Architectures

We have employed CNN-LSTM by Veltri et al. (2018), CNN-BiLSTM, CNN, BiLSTM, stacked LSTM, stacked BiLSTM, CNN-RNN, CNN-GRU to our curated dataset composed of both AMPs and decoy peptides. The results are shown in Table 1. Reduced performance compared to their original dataset consisting only of linear cationic peptides was observed upon application of Veltri et al. CNN-LSTM to our dataset. Deep learning model performances reduced drastically on a heterogenous dataset reflecting real world conditions. However, deep learning approaches in general provided satisfactory performance for AMP recognition problem. It is observed that metrics such as accuracy and recall is easier to improve whereas precision and MCC are harder. This may be due to the rich and diverse set of AMPs constituting our dataset which in turn may provide a variety of statistical patterns for the models and therefore makes it harder for the model to be unbiased towards any subset of AMP or any class (AMP or decoy).

These performance results indicate that sequence modelling architectures such as RNN and LSTMs are not powerful modellers on their own to the level that matches a sufficient in task performance. On the other hand, a simple CNN model can achieve good performance even so with highest precision. The causative reason for this drastic performance difference may be the short lengths of peptides unlike proteins. The short sequence lengths might not allow the RNN alike architectures to model the sequence with such amount of data whereas the hidden evolutionary signals contained in the sequences could be exploited by convolution operations. That being said, as the BiLSTM model which takes account of both directions of sequence modelling also yields near SOTA results, the combination with CNN with a dropout layer to prevent overfitting is utilized and it performed as the best model in all metrics except precision. Since the loss difference was much higher and there was a dropout layer in the model, it was chosen over raw CNN model to be used in production.

For practical purposes, we have choosen our final model based on precision and MCC values. Solely, because, MCC is a better performance indicator for imbalanced datasets such as ours and that precision enables us to compare False Positive (FP) rates across models which can reduce unnecessary false attempts in both computational

verification and wet-laboratory validation and thus reduce unwanted costs. As it can be seen in the table, CNN based models tend to perform better in terms of precision and MCC, and while not so much in accuracy and AUC. BiLSTM models tend to perform better in terms of recall. Although, there is a small precision performance difference between CNN and CNN-BiLSTM, combining CNN and BiLSTM with a dropout layer for generalization, provides best overall performance metrics for our practical goals as evident by the corresponding precision, MCC and loss values. Hence, CNN-BiLSTM model is chosen to be the production model with which AMP candidates are scanned in several data dumps and CDSs of organisms.

Table 1. Performance results of deep learning models using various metrics. Bold rows are the best performing model at that metric.

	Loss	**Accuracy**	**AUC**	**Precision**	**Recall**	**MCC**
Veltri et. al. (CNN-LSTM)	0.283	0.888	0.952	0.768	0.915	0.756
CNN-BiLSTM	**0.239**	**0.895**	**0.962**	0.774	**0.935**	**0.773**
CNN	0.283	0.889	0.953	**0.786**	0.881	0.749
BiLSTM	0.287	0.87	0.949	0.734	0.917	0.728
Stacked LSTM	0.693	0.461	0.50	0.335	0.60	0.001
Stacked BiLSTM	0.288	0.873	0.949	0.737	0.919	0.731
CNN-RNN	0.289	0.882	0.950	0.757	0.912	0.745
CNN-GRU	0.293	0.887	0.952	0.776	0.899	0.753

4.3 Uniprot Data Dump

The net charge and hydropathy distribution of the peptide fragments with a prediction score of 1 from Uniprot data dump are given in Figure 8 and Figure 9. Both the net charge and hydropathy score densities peak around value of 0 (peak hydropathy scores closer to -0.5), being and follow a Gaussian distribution which indicates that these features are not distinctly informative for the prediction task but provide utility in interpreting predicted AMPs with tail values. Hydropathy profiles indicate that peptide fragments in this data dump are largely hydrophilic. This also shows that the dataset is heterogenous and our model learns with variety.

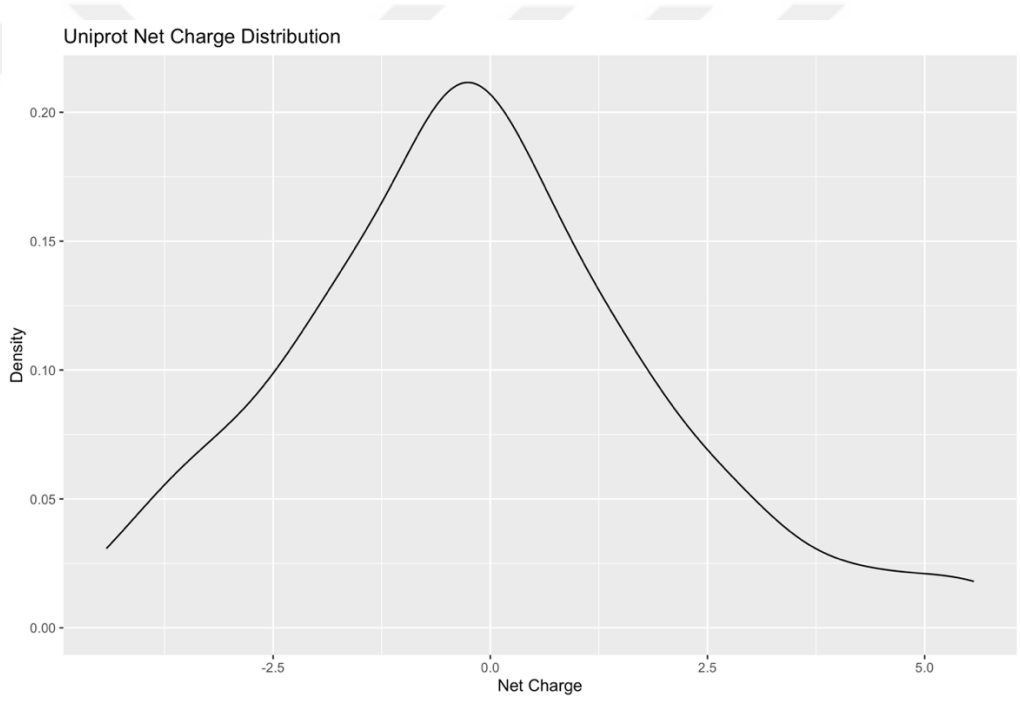


Figure 8. Net charge distribution of Uniprot peptide fragments with a prediction score of 1.

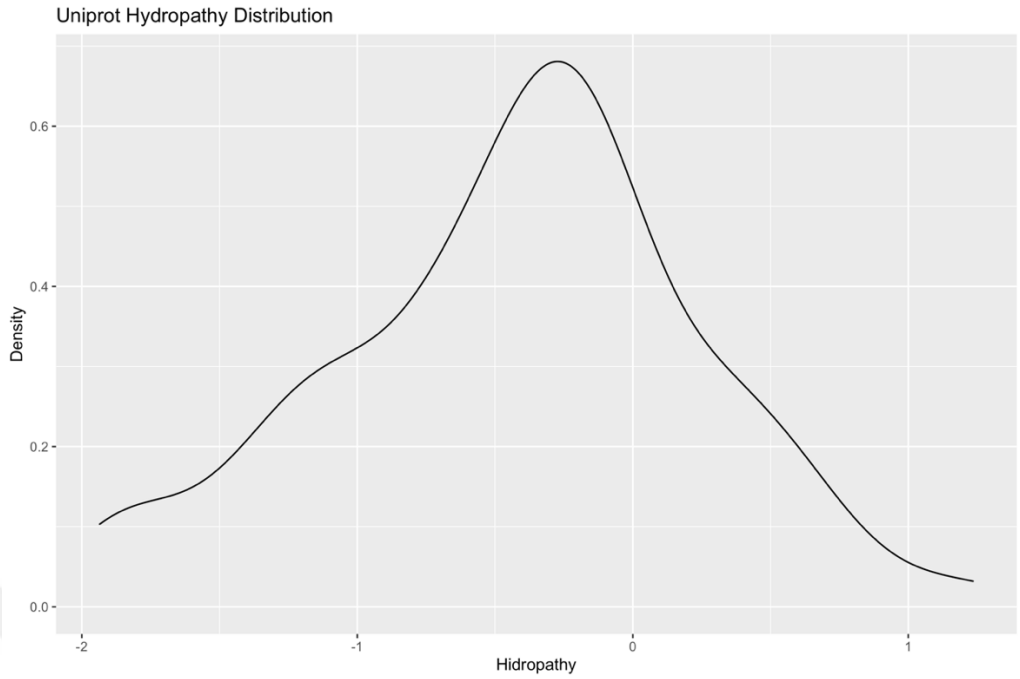


Figure 9. Hydropathy distribution of Uniprot peptide fragments with a prediction score of 1.

The net charge and hydropathy distribution of the peptide fragments with a prediction score of 1 from *B. subtilis* 15 and 30 window size scans are given in Figure 10, Figure 11, Figure 12, and Figure 13 respectively. Both the net charge and hydropathy score densities peak around value of 0 (peak hydropathy scores closer to -0.5) and follow a Gaussian-like distribution for 15 amino acid long peptides which indicates that these features are not distinctly informative for the prediction task but provide utility in interpreting predicted AMPs with tail values. With 30 amino acid long peptides, the distributions did not follow a Gaussian pattern. This also shows that the dataset is heterogenous and our model learns with variety.

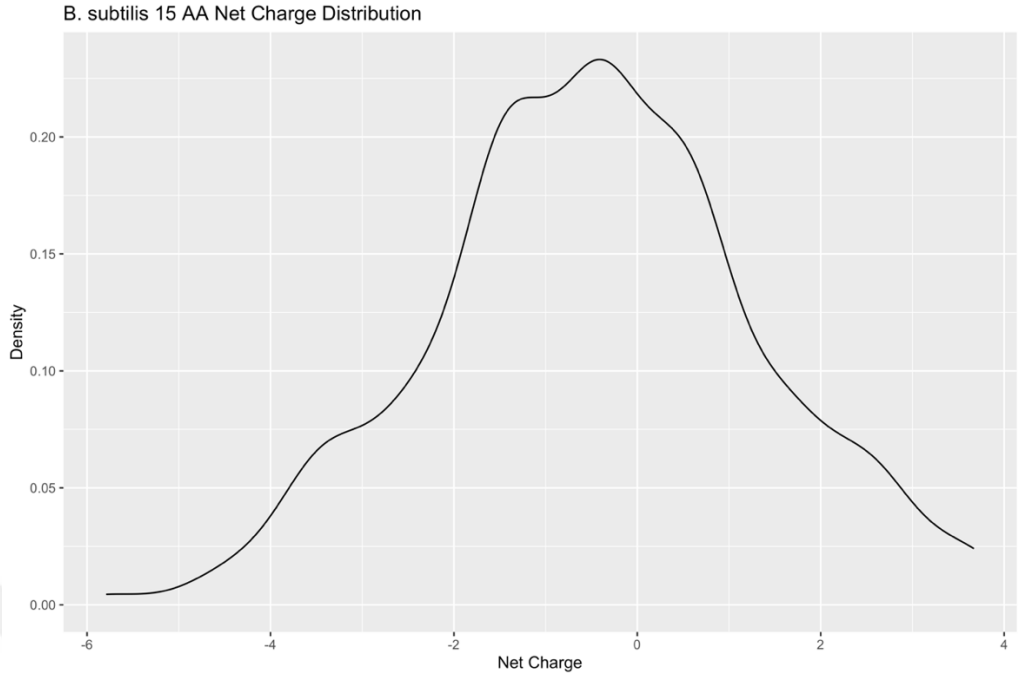


Figure 10. Net charge distribution of 15 amino acid long peptides from *Bacillus subtilis* proteome.

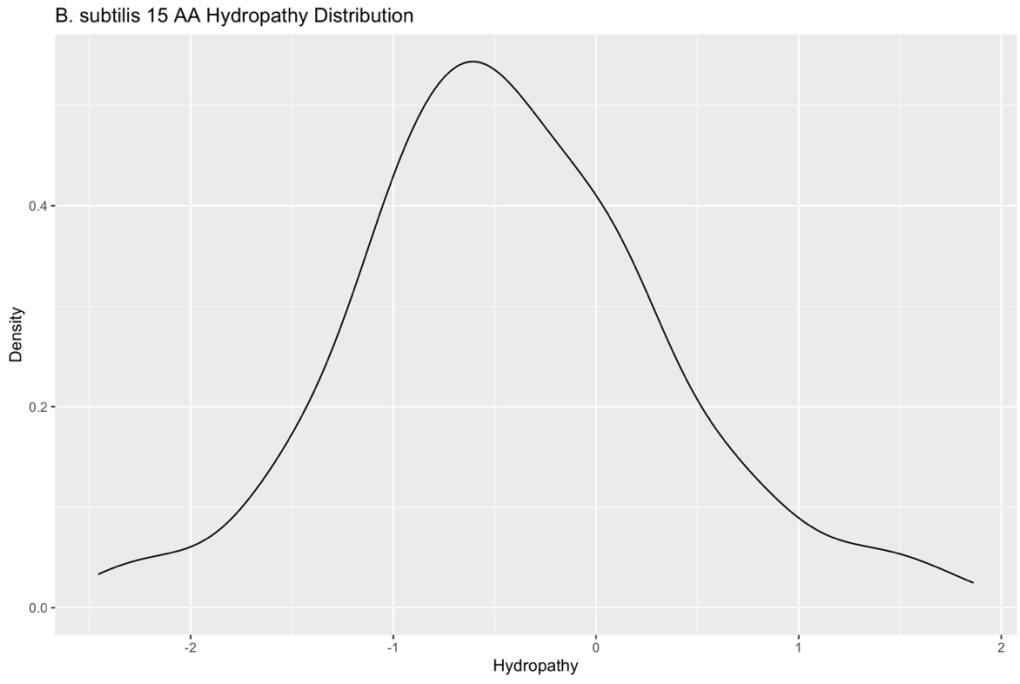


Figure 11. Hydropathy score distribution of 15 amino acid long peptides from *Bacillus subtilis* proteome.

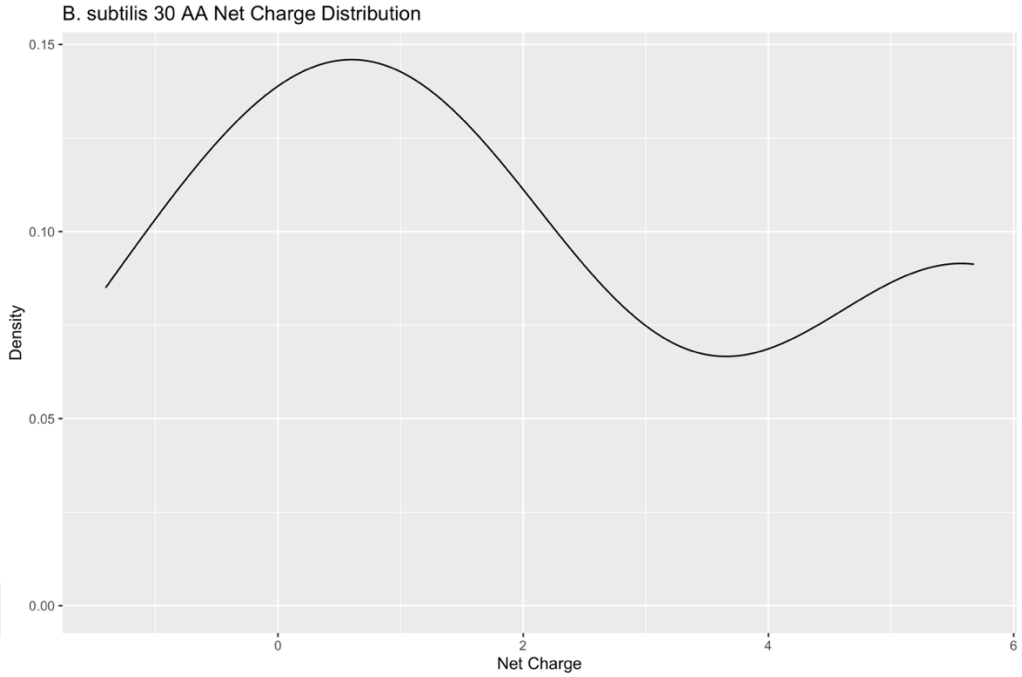


Figure 12. Net charge distribution of 30 amino acid long peptides from *Bacillus subtilis* proteome.

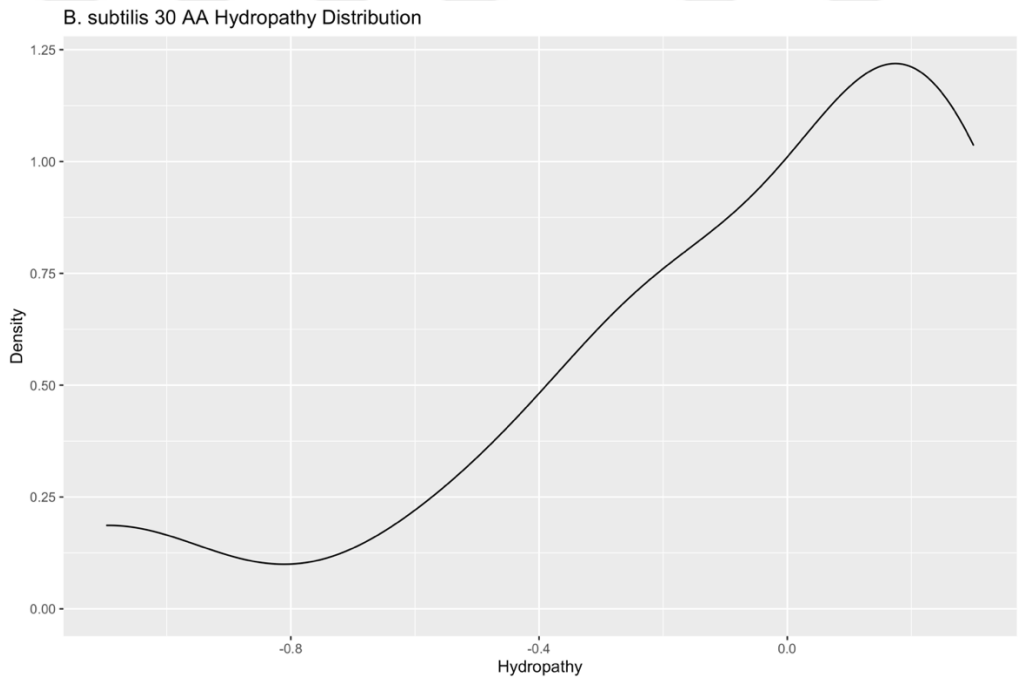


Figure 13. Hydropathy score distribution of 30 amino acid long peptides from *Bacillus subtilis* proteome.

The net charge and hydropathy distribution of the peptide fragments with a prediction score of 1 from *Trichoderma reesei* 20 and 30 window size scans are given in Figure 14, Figure 15, Figure 16, Figure 17 respectively. Both the net charge and hydropathy score densities peak around value of 0 (peak hydropathy scores closer to -0.5) and follow a Gaussian-like distribution which indicates that these features are not distinctly informative for the prediction task but provide utility in interpreting predicted AMPs with tail values. This also shows that the dataset is heterogenous and our model learns with variety.

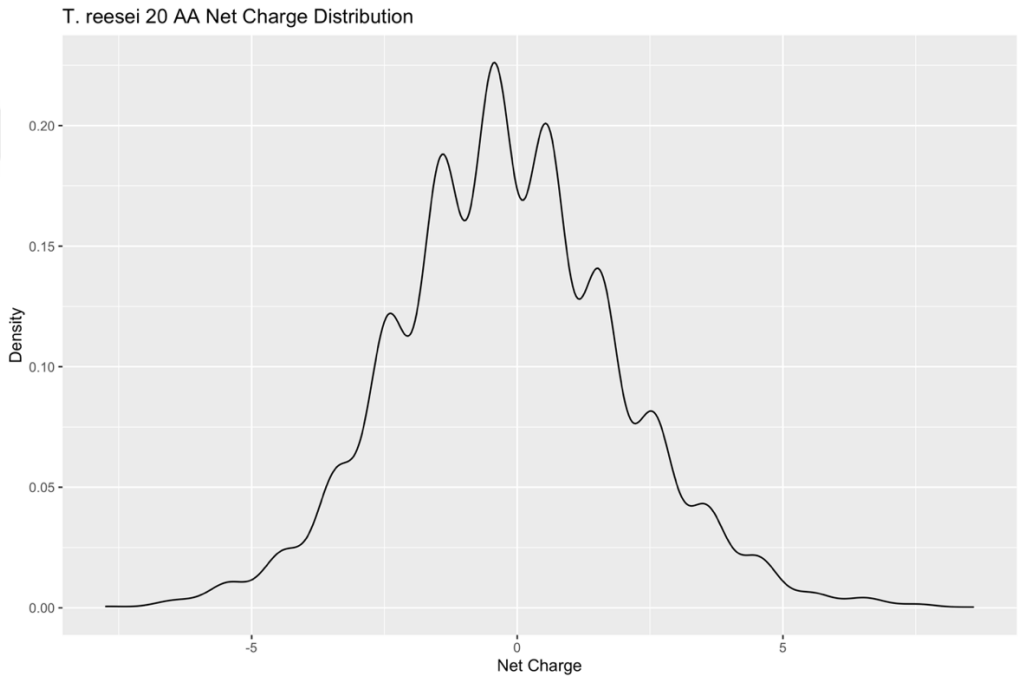


Figure 14. Net charge distribution of 20 amino acid long peptides from *Trichoderma reesei* proteome.

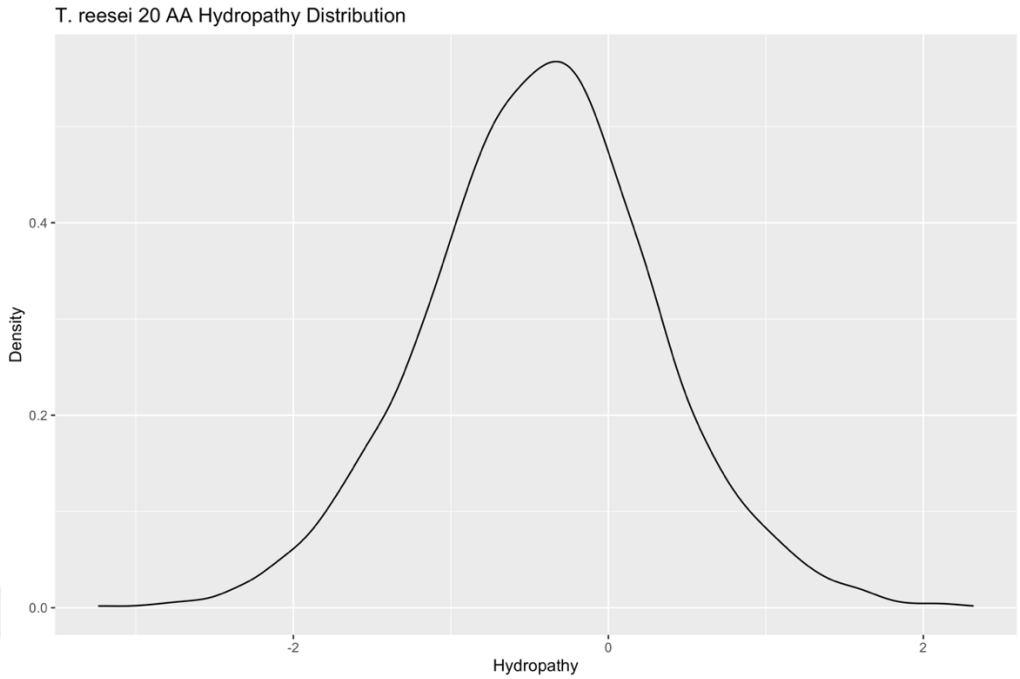


Figure 15. Hydropathy score distribution of 20 amino acid long peptides from *Trichoderma reesei* proteome.

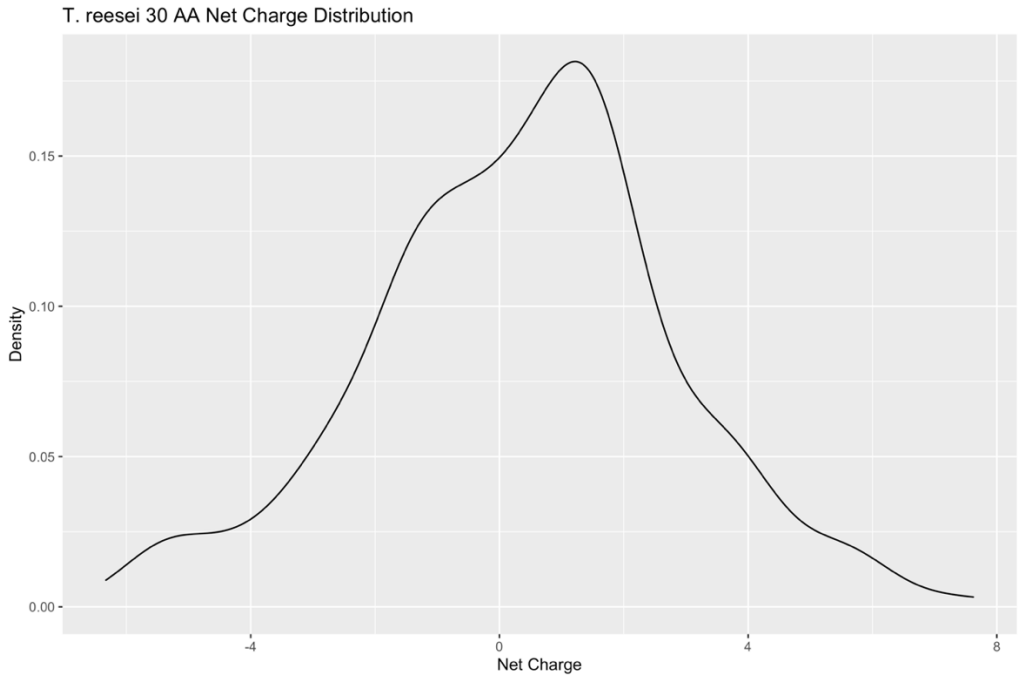


Figure 16. Net charge distribution of 30 amino acid long peptides from *Trichoderma reesei* proteome.

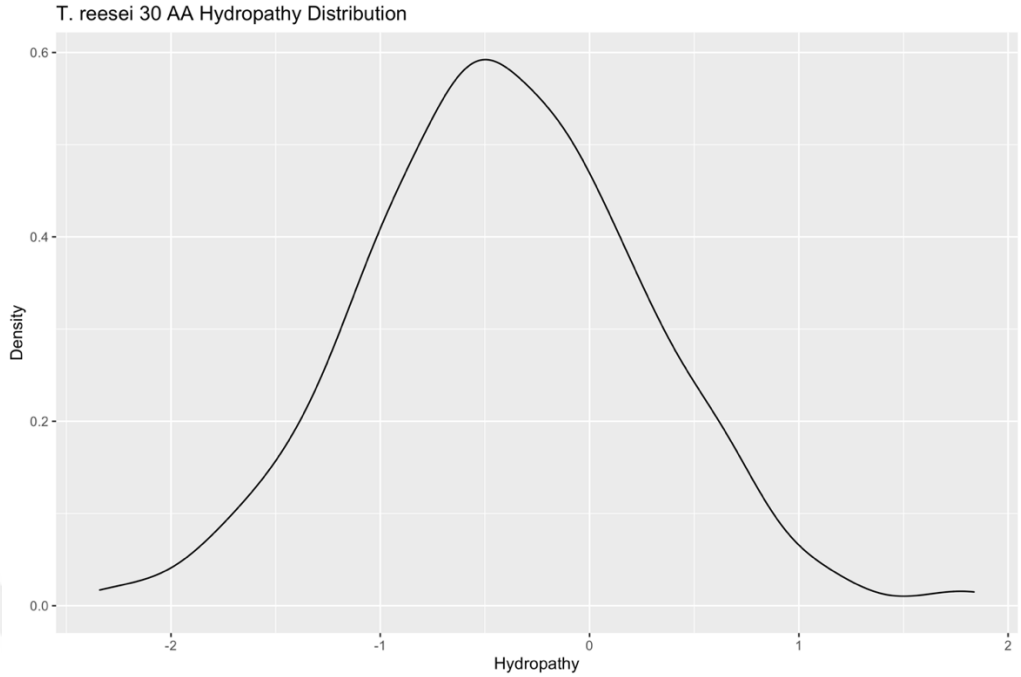


Figure 17. Net charge distribution of 30 amino acid long peptides from *Trichoderma reesei* proteome.

4.4 Prediction Results

CNN-BiLSTM model was applied on peptide fragments acquired from Uniref50 database, *B. subtilis* and *T. reesei* proteomes. Predicted outputs are annotated with their relevant biochemical and physico-chemical features to filter and rank them. Annotated results are also used to further extrapolate according to which features the model classifies and misclassifies the peptides.

4.4.1 UniRef50 prediction results

The resulting number of peptide fragments with size of 20 classified as AMP by our model was extensively high. That is, the number of peptides with classification score between 0.50 and 1 is around 67.000. For this reason, only the peptides with a classification score of 1 are subjected for further inspection with the apriori assumption of these being a false negative are of low probability. After the inspection, 310 peptides are analyzed according to the filtering and prioritization procedure outlined in the methods section. Top 10 ranked candidate AMPs, sorted by their related hydropathy scores, are shown in Table 2 along with their biochemical and physicochemical features used to filter and rank them. These candidate AMP sequences then aligned to non-redundant protein database using BlastP to forge biologically framed hypothesis that supports their candidate potentials.

First, random, 10 peptides predicted as AMP with a prediction score of 1, are observed to be both positively charged and negatively charged. They also observed to be both hydrophilic and hydrophobic. These results contribute to the preliminary observation that simple features cannot be used to discriminate between AMPs and decoys.

Table 2. First 10 sequence from Uniprot with a prediction score of 1 and their features.

Sequence	Hydropathy	Net Charge	Helix Fraction	Turn Fraction	Sheet Fraction	Instability index
ECQRKLKAVYERRRHISLEP	-1.305	2.674149	0.25	0.1	0.3	110.18
GVITVNRHQVRLFTDSQIAL	0.25	0.594476	0.4	0.15	0.15	34.62
NMKELRFLHVDCPVGSEEEV	-0.4	-3.42803	0.3	0.2	0.35	68.68
ENTLKVELEFNKMLPNSVC	-0.29	-1.35987	0.35	0.25	0.35	46.375
TEERALEIAYENKIGKYYMA	-0.735	-1.79968	0.3	0.1	0.45	78.95
VVEKTVILSQSNQRFEKLTS	-0.045	0.520356	0.35	0.2	0.25	46.485
IYRICPSKISKRELEDLYFA	-0.245	0.525565	0.4	0.15	0.25	41.99
LALSEARARAPSVLVFDELD	0.37	-2.43819	0.35	0.15	0.5	44.56
LIDLPILNQNSLNLILTIPG	0.91	-1.4421	0.5	0.35	0.3	33.18
AEDSVRQDTEDCSRYIEAIS	-0.93	-4.41456	0.2	0.15	0.25	87.595

As can be seen in Table 2, the top BLAST results from of the 10 random predicted Uniprot AMPs are tend to hit mostly to enzymes, receptors or transcriptional regulators. Enzymes such as polymerases, histidine kinase and ATPase align with chargeness and hydropathicity properties of AMP mechanisms which enable the AMPs to either bind to their target or enact their effects.

Table 3. BlastP results of first 10 sequence from Uniprot.

Sequence	BlastP First Hit	Secondary Hits
ECQRKLKAVYERRRHISLEP	phosphotransferase [Trypanosoma cruzi]	
GVITVNRHQVRLFTDSQIAL	histidine kinase [uncultered bacterium]	
NMKELRFLHVDCPVGSEEEV	toll/interleukin-1 receptor (TIR) domain-containing protein [Artemisia annua]	
ENTLKVELEFNKMLPNSVC	CLUMA_CG008453, isoform A [Clunio marinus]	
TEERALEIAYENKIGKYYMA	RNA-dependent RNA polymerase [Lihan tick virus]	
VVEKTVILSQSNQRFEKLTS	PAS domain-containing protein [Geminocystis herdmanii]	sigma-54-dependent Fis family transcriptional regulator [Labibibaculum sp.]
IYRICPSKISKRELEDLYFA	uncharacterized protein LOC116778816 isoform X1 [Danaus plexippus plexippus]	
LALSEARARAPSVLVFDELD	conserved unknown protein [Ectocarpus siliculosus]	AAA family ATPase [Paracoccus mutanolyticus]
LIDLPILNQNSLNLILTIPG	TonB-dependent receptor [Edaphobacter aggregans]	
AEDSVRQDTEDCSRYIEAIS	hypothetical protein A8H39_18615 [Paraburkholderia fungorum]	NERD domain-containing protein [Paraburkholderia fungorum]

Table 4. First 10 sequence and their features sorted by net charge from Uniprot.

Sequence	Hydropathy	Net Charge	Helix Fraction	Turn Fraction	Sheet Fraction	Instability index
KARLNCIRGFLLEKIPHKRVK	-0.46	5.540599	0.35	0.1	0.25	29.075
KKKIATKNISFCVKKGEIL	-0.38	5.516531	0.3	0.15	0.15	33.26
YKLNYWIIKTNEKRRTRKEQ	-1.935	4.544585	0.3	0.1	0.15	42.88
SNRKCSKVFNFLFGGRRNNN	-1.115	4.223638	0.25	0.45	0.05	38.72
KIIQRGAPKNKIEKIYSFLL	-0.225	3.546021	0.4	0.2	0.2	31.715
SKGTVQADFWRFRQFNAKSK	-1.105	3.246108	0.25	0.2	0.1	-1.145
ECQRKLKAVYERRRHISLEP	-1.305	2.674149	0.25	0.1	0.3	110.18
EIGGKMDRKQINAALRKFLL	-0.34	2.660281	0.3	0.15	0.35	55.035
ATGRWWVKDNNLYVRARLLN	-0.61	2.603161	0.4	0.2	0.25	28.21
LLEWWKVNALRYPVLAKMAR	0.12	2.551007	0.45	0.1	0.45	41.87

Table 5. BlastP results of 10 sequence sorted by net charge.

Sequence	BlastP First Hit	Secondary Hits
KARLNCIRGFLLEKIPHKRVK	polyphosphate kinase 2, internal deletion [Myxococcus xanthus DK 1622]	
KKKIATKNISFCVKKGEIL	PREDICTED: ATP-binding casette sub-family A member 10-like [Chaetura pelagica]	
YKLNYWIIKTNEKRRTRKEQ	hypothetical protein [uncultured archaeon]	LysM peptidoglycan-binding domain-containing protein [Halothermothrix orenii]
SNRKCSKVFNFLFGGRRNNN	hypothetical protein PRIOE2DRAFT_16316 [Piromyces sp. E2]	
KIIQRGAPKNKIEKIYSFLL	hypothetical protein [Camplylobacter jejuni]	
SKGTVQADFWRFRQFNAKSK	serine protease [Candidatus Tenderia elecrophaga]	
ECQRKLKAVYERRRHISLEP	phosphotransferase [Trypanosoma cruzi]	
EIGGKMDRKQINAALRKFLL	phage major capsid protein [Dialister succinatiphilus]	
ATGRWWVKDNNLYVRARLLN	hypothetical protein CVV45_11955 [Spriochaetae bacterium HGW-Spriochaetae-10]	aminotransferase class I/II-fold pyridoxal phosphate-dependent enzyme [Paraferrimonas sp. SM1919] (partly similar)
LLEWWKVNALRYPVLAKMAR	zinc finger BED domain-containing protein RICESLEEPER 2 [Brachypodium distachyon]	

Table 6. First 10 sequence and their features sorted by hydropathy from Uniprot.

Sequence	Hydropathy	Net Charge	Helix Fraction	Turn Fraction	Sheet Fraction	Instability index
FEEFTRHLYLCCILVLEFLL	1.235	-2.4541	0.6	0	0.45	450.06
LIDLPILNQNSNLNLILTIPG	0.91	-1.4421	0.5	0.35	0.3	33.18
LSLLCSWVQALLAGRGDWKL	0.675	0.530865	0.45	0.2	0.4	-3.235
WVHQHAVLNHISNLTLIQVV	0.65	-0.33267	0.5	0.15	0.2	23.455
ANGECKAIVWLDENVLLIE	0.625	-3.41331	0.45	0.15	0.45	44.08
WEASIIASYWLSHLAATNAA	0.55	-1.40731	0.35	0.2	0.45	2.97
SIYDVYSSLFAWVNLQENII	0.535	2.75022	0.55	0.25	0.2	80.855
PVVIGHKFASVTVLDSEPAD	0.4	-2.06127	0.35	0.25	0.2	61
NAKSCRVVSSIFTILRNMEV	0.395	1.531514	0.35	0.25	0.2	60.86
LALSEARARAPSVLVFDELD	0.37	2.43819	0.35	0.15	0.5	44.56

Table 7. BlastP results of 10 sequence sorted by net charge by hydropathy.

Sequence	BlastP First Hit	Secondary Hits
FEEFTRHLYLCCILVLEFLL	unnamed protein product [Oncorhynchus mykiss]	Partially LysR family transcriptional regulator [Thermoanaerobacter uzonensis]
LIDLPILNQNSNLNLILTIPG	TonB-dependent receptor [Edaphobacter aggregans]	
LSLLCSWVQALLAGRGDWKL	hypothetical protein [Acidipila sp. 4G-K13]	Partial similarity: potassium-transporting ATPase subunit KdpB [Aquaspirillum sp. LM1]
WVHQHAVLNHISNLTLIQVV	reverse transcriptase zinc-binding domain-containing protein [Artemisia annua]	
ANGECKAIVWLDENVLLIE	hypothetical protein DRQ44_00025 [Gammaproteobacteria bacterium]	Partial similarity: Uma2 family endonuclease [Leptolyngbya sp. DLM2.Bin27]
WEASIIASYWLSHLAATNAA	ankyrin reppeat containing protein [Theileria orientalis]	Partial similarity: putative methyltransferase family protein [Theileria parva strain Muguga]
SIYDVYSSLFAWVNLQENII	ABC transporter permase [Bacteroidetes bacterium]	
PVVIGHKFASVTVLDSEPAD	potassium transporter Kup [Deltaproteobacteria bacterium]	
NAKSCRVVSSIFTILRNMEV	uncharacterized protein LOC111612615 [Centruroides sculpturatus]	Partial similarity: serine/teronine protein kinase, putative [Plasmodium ovale wallikeri]
LALSEARARAPSVLVFDELD	conserved unknown protein [Ectocarpus siliculosus]	Partial similarity: AAA family ATPase [*Paracoccus mutanolyticus]*

4.4.2 *Bacillus subtilis* peptides prediction results

The resulting number of peptide fragments with size of 15 and size 30 from *B. subtilis.* CDS regions classified as AMP by our model is extensively high. For this reason, only the peptides with a classification score of 1 are subjected for further inspection with the *apriori* assumption of these being a false negative are of low probability. These are in number 198 when 15 AA length window size is used, 12 when 30 AA length window size is used. The predicted AMP sequences and their features resulting from the 30 window size scans are given in Table 8 and Table 9. After the inspection, 310 peptides are analyzed according to the filtering and prioritization procedure outlined in the methods section. Top 10 ranked candidate AMPs, sorted by their related hydropathy scores, are shown in Table 10 along with their biochemical and physicochemical features used to filter and rank them. These candidate AMP sequences then aligned to non-redundant protein database using BlastP to forge biologically framed hypothesis that supports their candidate potentials.

Table 8. First 10 peptides (30 aminoacid long peptides) and their physicochemical, structural properties predicted from *Bacillus subtilis* proteome with prediction score 1.

Sequence	Hydropathy	Net Charge	Helix Fraction	Turn Fraction	Sheet Fraction	Instability index
AIWAHRIAHALYKRKFYFLARLISQVSRFF	0.28333333	5.67093234	0.46666667	0.06666667	0.26666667	31.057
VVAPAKVLNDFAKNHEALEIKAGVIEGKVS	0.21333333	-0.4464363	0.33333333	0.2	0.33333333	29.74
TEYLAVRISMEWALIKPYLPLTNERSKKTI	-0,2233333	1.19917648	0.36666667	0.16666667	0.33333333	16.78
EFSVALSTLWQLISRTNKYIDETAPWVLAK	0.06	-0.3385732	0.4	0.16666667	0.3	29.4933333
EKGLAHNIKSAKRKIERVQPEVWDVLESVI	-0.51	0.69776402	0.33333333	0.16666667	0.26666667	65.0333333
IWAHRIAHALYKRKFYFLARLISQVSRFFT	0.2	5.62058291	0.46666667	0.06666667	0.23333333	31.057
RLARNFEHIRFPTRYSNEEASSIIQQFEDQ	-1.0966667	-1.4035972	0.26666667	0.2	0.23333333	52.4233333
IKPKHSAFYETALHTLLTELQVRHIIITGI	0.3	0.66203182	0.4	0.1	0.26666667	40.3436667
EYLAVRISMEWALIKPYLPLTNERSKKTIS	-0.2266667	1.65706474	0.36666667	0.2	0.33333333	10.36
HAIWAHRIAHALYKRKFYFLARLISQVSRF	0.08333333	5.65720935	0.43333333	0.06666667	0.26666667	31.057
AHRIAHALYKRKFYFLARLISQVSRFFTGI	0.21666667	5.67093234	0.43333333	0.1	0.23333333	30.407
AEALKDIIPSSKPKVITIKEIQRVVGQQFN	-0.1733333	1.60044282	0.33333333	0.2	0.16666667	18.04

Table 9. BlastP results for the first 10 peptides (30 aminoacid long peptides) predicted from *Bacillus subtilis* proteome with prediction score 1.

Sequence	First hit of BlastP	Secondary and other hits
AIWAHRIAHALYKRKFYFLARLISQVSRFF	serine O-acetyltransferase [*Bacillus* sp. RHFS10]	
VVAPAKVLNDFAKNHEALEIKAGVIEGKVS	50S ribosomal protein L10 [*Bacillus tequilensis*]	
TEYLAVRISMEWALIKPYLPLTNERSKKTI	DUF2309 family protein [*Xanthomonas citri pv. citri]*	YbcC family protein [*Bacillus subtilis*]
EFSVALSTLWQLISRTNKYIDETAPWVLAK	class I tRNA ligase family protein [*Escherichia coli]*	methionyl-tRNA synthetase [*Bacillus* sp. 5B6]
EKGLAHNIKSAKRKIERVQPEVWDVLESVI	hypothetical protein [*Klebsiella pneumoniae*]	DNA-directed RNA polymerase subunit beta' [*Pseudomonas* sp. PGSB 8459]
IWAHRIAHALYKRKFYFLARLISQVSRFFT	serine O-acetyltransferase [*Bacillu*s sp. RHFS10]	
RLARNFEHIRFPTRYSNEEASSIIQQFEDQ	protein arginine kinase [*Bacillus subtilis*]	
IKPKHSAFYETALHTLLTELQVRHIIITGI	MULTISPECIES: cysteine hydrolase [Bacillaceae]	isochorismatase family protein [*Bacillus subtilis*]
EYLAVRISMEWALIKPYLPLTNERSKKTIS	DUF2309 family protein [*Xanthomonas citri* pv. citri]	YbcC family protein [*Bacillus subtilis*]
HAIWAHRIAHALYKRKFYFLARLISQVSRF	serine O-acetyltransferase [*Bacillus* sp. RHFS10]	
AHRIAHALYKRKFYFLARLISQVSRFFTGI	serine O-acetyltransferase [Bacillus sp. RHFS10]	
AEALKDIIPSSKPKVITIKEIQRVVGQQFN	chromosomal replication initiation protein [*Bacillus subtilis* subsp. subtilis str. SC-8]	

Table 10. First 10 peptides (15 aminoacid long peptides) and their physicochemical, structural properties predicted from *Bacillus subtilis* proteome with prediction score 1.

Sequence	Hydropathy	Net Charge	Helix Fraction	Turn Fraction	Sheet Fraction	Instability index
RDKSVICVVQDPKDV	-0.3266667	-0.4707474	0.33333333	0.13333333	0	11.4866667
KKQTNHRVEVIFTKE	-1.2733333	1.59087142	0.26666667	0.06666667	0.13333333	15.0733333
YHGLKDVEQRYRQRY	-2.06	1.58554883	0.33333333	0.06666667	0.13333333	9.34
YVRTKDHPLIEELEK	-1.12	-1.4096538	0.33333333	0.06666667	0.33333333	58.44
KYHGLKDVEQRYRQR	-2.2333333	2.58554883	0.26666667	0.06666667	0.13333333	14.3666667
VAEMKVYARQERVEL	-0.42	0.4785483	0.33333333	0	0.46666667	40.1133333
TASEKKKIEYLDKTY	-1.3866667	0.19604685	0.26666667	0.06666667	0.26666667	40.1466667
RLMCCLKYENDEYET	-0.98	-2.495287	0.26666667	0.06666667	0.4	6.46
QNGWVRLASLDPEGC	-0.4333333	-1.4655084	0.26666667	0.33333333	0.26666667	13.4666667
KGLHHLVWEIVDNSI	0.14	-1.3702326	0.46666667	0.2	0.2	0.83333333

Table 11. BlastP results for the first 10 peptides (15 aminoacid long peptides) predicted from *Bacillus subtilis* proteome with prediction score 1.

Sequence	First hit of BlastP	Secondary and other hits
RDKSVICVVQDPKDV	recombination protein RecR [*Bacillus subtilis*]	
KKQTNHRVEVIFTKE	Crystal structure of Rae1 (YacP) from *Bacillus subtilis* (W164L mutant) [*Bacillus subtilis* subsp. subtilis str. 168]	ribosome-dependent mRNA decay endonuclease Rae1/YacP *[Bacillus mojavensis]*
YHGLKDVEQRYRQRY	lysine--tRNA ligase [*Oceanobacillus senegalensis*]	
YVRTKDHPLIEELEK	nucleoside triphosphate pyrophosphohydrolase [*Bacillus atrophaeus*]	
KYHGLKDVEQRYRQR	lysine--tRNA ligase [*Oceanobacillus senegalensis*]	lysyl-tRNA synthetase *[Exiguobacterium oxidotolerans]*
VAEMKVYARQERVEL	transcription-repair coupling factor [*Bacillus subtilis*]	
TASEKKKIEYLDKTY	iron ABC transporter substrate-binding protein FeuA [*Bacillus subtilis*]	
RLMCCLKYENDEYET	stage 0 sporulation family protein [*Bacillus* sp. REN3]	Cell fate regulator YaaT, PSP1 superfamily (controls sporulation, competence, biofilm development) *[Paenisporosarcina quisquiliarum]*
QNGWVRLASLDPEGC	YbcC family protein [*Bacillus subtilis*]	DUF2309 family protein *[Bacillus subtilis]*
KGLHHLVWEIVDNSI	TPA: DNA gyrase [*Lysinibacillus* sp.]	type IIA DNA topoisomerase subunit B [Maize bushy stunt phytoplasma]

Table 12. First 10 peptides (15 aminoacid long peptides) with the highest net charge and their physicochemical, structural properties predicted from *Bacillus subtilis* proteome with prediction score 1.

Sequence	Hydropathy	Net Charge	Helix Fraction	Turn Fraction	Sheet Fraction	Instability index
ERRIERRVRKITSEK	-1.94	3.66450616	0.2	0.06666667	0.2	122.18
KMVRRHPHVFKDVKV	-0.7133333	3.62358522	0.33333333	0.06666667	0.06666667	62.5733333
KLGRTSAQRKAMLRD	1.14	3.55408212	0.13333333	0.13333333	0.33333333	16.5133333
RCERCGRPHSVIRKF	-0.9933333	3.54358727	0.2	0.2	0.06666667	22.1733333
CERCGRPHSVIRKFK	-0.9533333	3.54110679	0.2	0.2	0.06666667	12.1533333
EAEIWGRRIRAFRKL	-0.7333333	2.66591611	0.33333333	0.06666667	0.33333333	47.7866667
KYHGLKDVEQRYRQR	-2.2333333	2.58554888	0.26666667	0.06666667	0.13333333	14.3666667
KLISVSLPRVRDFRG	-0.08	2.5553184	0.4	0.26666667	0.13333333	0.8
GQKQRVAIAGVIAAR	0.26	2.55489704	0.26666667	0.13333333	0.26666667	25.0933333
QKQRVAIAGVIAARP	0.18	2.55489704	0.26666667	0.13333333	0.26666667	37.9333333

Table 13. BlastP results for the first 10 peptides with the highest net charge (15 aminoacid long peptides) predicted from *Bacillus subtilis* proteome with prediction score 1.

Sequence	First hit of BlastP	Secondary and other hits
ERRIERRVRKITSEK	Crystal structure of Rae1 (YacP) from *Bacillus subtilis* (W164L mutant) [*Bacillus subtilis* subsp. subtilis str. 168]	ribosome-dependent mRNA decay endonuclease Rae1/YacP [*Bacillus mojavensis*]
KMVRRHPHVFKDVKV	nucleoside triphosphate pyrophosphohydrolase [*Bacillus velezensis*]	
KLGRTSAQRKAMLRD	50S ribosomal protein L17 [*Acinetobacter baumannii*]	
RCERCGRPHSVIRKF	30S ribosomal protein S14 [*Escherichia coli*]	type Z 30S ribosomal protein S14 [*Bacillus licheniformis*]
CERCGRPHSVIRKFK	30S ribosomal protein S14 [*Escherichia coli*]	type Z 30S ribosomal protein S14 [*Bacillus licheniformis*]
EAEIWGRRIRAFRKL	MULTISPECIES: helix-turn-helix transcriptional regulator [*Bacillus*]	
KYHGLKDVEQRYRQR	lysine--tRNA ligase [*Oceanobacillus senegalensis*]	lysyl-tRNA synthetase [*Exiguobacterium oxidotolerans*]
KLISVSLPRVRDFRG	50S ribosomal protein L5 [Candidatus Riflebacteria bacterium]	
QKQRVAIAGVIAARP	energy-coupling factor ABC transporter ATP-binding protein [*Bacillus siamensis*]	
GQKQRVAIAGVIAAR	energy-coupling factor ABC transporter ATP-binding protein [*Bacillus siamensis*]	

Table 14. First 10 peptides (15 aminoacid long peptides) with the highest hydropathy and their physicochemical, structural properties predicted from *Bacillus subtilis* proteome with prediction score 1.

Sequence	Hydropathy	Net Charge	Helix Fraction	Turn Fraction	Sheet Fraction	Instability index
ALFVLHIFMPLFAWG	1.86	-0.3555713	0.6	0.13333333	0.4	61.0066667
IITLCMVLLPQSSSI	1.68666667	-0.4670505	0.46666667	0.26666667	0.26666667	111.72
LKGLNVIAPLGFIFA	1.6	0.55494727	0.53333333	0.26666667	0.33333333	10.8533333
ANAAAIAIAEIVAGS	1.47333333	-1.3910769	0.26666667	0.2	0.53333333	16.5133333
AVIVGVIFIQQAVRK	1.39333333	1.60527159	0.53333333	0.06666667	0.13333333	10.8533333
WANWIAIVSVIPVEA	1.35333333	-1.4414264	0.53333333	0.2	0.26666667	-2.34
PIGAVLVINGEIIAR	1.32666667	-0.0975763	0.46666667	0.26666667	0.26666667	1.46666667
GWANWIAIVSVIPVE	1.20666667	-1.4393339	0.53333333	0.26666667	0.2	5.88666667
LLLPAFVSGYFPWCD	1.02666667	-1.4678403	0.53333333	0.26666667	0.26666667	53.1133333
LGNYEWAVPWIFAFI	0.94	-1.443932	0.6	0.2	0.26666667	-20.493333

Table 15. BlastP results for the first 10 peptides with the highest net hydropathy (15 aminoacid long peptides) predicted from *Bacillus subtilis* proteome with prediction score 1.

Sequence	First hit of BlastP	Secondary and other hits
ALFVLHIFMPLFAWG	hypothetical protein AN935_00830 [*Bacillus inaquosorum*]	putative sodium dependent transporter [*Bacillus inaquosorum* KCTC 13429]
IITLCMVLLPQSSSI	iron ABC transporter permease [*Bacillus subtilis*]	
LKGLNVIAPLGFIFA	amino acid permease [*Bacillus subtilis* subsp. subtilis str. RO-NN-1]	putative amino acid-proton symporter YbeC [*Bacillus subtilis* PY79]
ANAAAIAIAEIVAGS	D-alanyl-D-alanine carboxypeptidase [*Bacillus subtilis*]	
AVIVGVIFIQQAVRK	preprotein translocase subunit SecY [*Bacillus cereus* group sp. N6]	
WANWIAIVSVIPVEA	putative amino acid-polyamine-organocation superfamily protein [*Xanthomonas citri* pv. fuscans]	amino-acid permease protein [*Mizugakiibacter sediminis*]
PIGAVLVINGEIIAR	tRNA specific adenosine deaminase [*Bacillus* sp. LM 4-2]	
GWANWIAIVSVIPVE	putative amino acid-polyamine-organocation superfamily protein [*Xanthomonas citri* pv. fuscans]	MULTISPECIES: APC family permease [*Geobacillus*]
LLLPAFVSGYFPWCD	hypothetical protein B4067_0175 [*Bacillus subtilis* subsp. subtilis]	GNAT family N-acetyltransferase [*Bacillus subtilis*]
LNGYEWAVPWIFAFI	hypothetical protein AN935_00830 [*Bacillus inaquosorum*]	putative sodium dependent transporter [*Bacillus inaquosorum* KCTC 13429]

4.4.3 *Trichoderma reesei* peptides prediction results

The resulting number of peptide fragments with size of 20 and 30 from *T. reseei.* CDS regions classified as AMP by our model is extensively high. For this reason, only the peptides with a classification score of 1 are subjected for further inspection with the apriori assumption of these being a false negative are of low probability. These are in number 3897 when 20 AA length window size is used, 251 when 30 AA length window size is used. The predicted AMP sequences and their features resulting from the 30 window size scans are given in Table 16 and Table 17. After the inspection, 250 peptides are analyzed according to the filtering and prioritization procedure outlined in the methods section. Top 10 ranked candidate AMPs, sorted by their related hydropathy scores, are shown in Table 18 along with their biochemical and physicochemical features used to filter and rank them. These candidate AMP sequences then aligned to non-redundant protein database using BlastP to forge biologically framed hypothesis that supports their candidate potentials.

Table 16. First 10 peptides (20 aminoacid long peptides) and their physicochemical, structural properties predicted from *Trichoderma reesei* proteome with prediction score 1.

Sequence	Hydropathy	Net Charge	Helix Fraction	Turn Fraction	Sheet Fraction	Instability index
KASEREADERRARGRSRFRS	-2.2	3.55858021	0.05	0.2	0.3	92.85
DADVSVRVAAVSVIDSLRAA	0.82	-1.4412579	0.35	0.15	0.3	24.515
ERCLQLDSFQQLRQDTSIYE	-0.97	-2.3555516	0.3	0.1	0.25	72.295
LRSFDTTCELSARGNLDEAP	-0.56	-2.4639662	0.2	0.25	0.35	40.895
WGELEERKTSEREKRRRAAL	-1.97	1.5579199	0.15	0.2	0.45	100.055
YSLDDEVLIDACWAISYLSD	0.365	-5.4678857	0.45	0.15	0.3	35.795
SLDDEVLIDACWAISYLSDG	0.41	-5.770874	0.4	0.2	0.3	35.795
SYPRHLILAIIAKLPSYIKL	0.695	2.27980726	0.5	0.2	0.3	37.685
DWIQHHINDIIANPGKLLDI	-0.115	-2.3704604	0.4	0.2	0.15	21.54
QGRAGRVQAGKCFKLFTQNL	-0.46	3.52788808	0.25	0.2	0.2	-10.275

Table 17. BlastP results for the first 10 peptides (20 aminoacid long peptides) predicted from *Trichoderma reesei* proteome with prediction score 1.

Sequence	First hit of BlastP	Secondary and other hits
KASEREADERRARGRSRFRS	RNA polymerase III RPC4 [*Trichoderma parareesei*]	
DADVSVRVAAVSVIDSLRAA	RNA polymerase III RPC4 [*Trichoderma parareesei*]	
ERCLQLDSFQQLRQDTSIYE	uncharacterized protein TRIREDRAFT_73501 [T. *reesei* QM6a]	STAG-domain-containing protein [*Trichoderma citrinoviride]*
LRSFDTTCELSARGNLDEAP	hypothetical protein A9Z42_0044760 [T. *parareesei*]	STAG-domain-containing protein [*Trichoderma citrinoviride*]
WGELEERKTSEREKRRRAAL	hypothetical protein A9Z42_0053800 [T. *parareesei]*	TPA: glutamate--tRNA ligase [*Calditthrix abyssi*]
YSLDDEVLIDACWAISYLSD	hypothetical protein L249_4871 *[Ophiocordyceps polyrhachis-furcata* BCC 54312]	putative ATP-dependent RNA helicase ucp12 [*Tolypocladium ophioglossoides*]
SLDDEVLIDACWAISYLSDG	hypothetical protein L249_4871 [*Ophiocordyceps polyrhachis-furcata*]	putative ATP-dependent RNA helicase ucp12 [*Tolypocladium ophioglossoides*]
SYPRHLILAIIAKLPSYIKL	uncharacterized protein TRIREDRAFT_53044 [T. *reesei* QM6a]	P-loop containing nucleoside triphosphate hydrolase protein [*Trichoderma citrinoviride*]
DWIQHHINDIIANPGKLLDI	uncharacterized protein TRIREDRAFT_53044 [T, *reesei* QM6a]	P-loop containing nucleoside triphosphate hydrolase protein [T. *citrinoviride*]
QGRAGRVQAGKCFKLFTQNL	uncharacterized protein TRIREDRAFT_53044 [T. *reesei* QM6a]	dead deah box helicase [*T. arundinaceum*]
KASEREADERRARGRSRFRS	RNA polymerase III RPC4 [*Trichoderma parareesei*]	
DADVSVRVAAVSVIDSLRAA	RNA polymerase III RPC4 [*Trichoderma parareesei*]	

Table 18. First 10 peptides with the highest net charge (20 aminoacid long peptides) and their physicochemical, structural properties predicted from *Trichoderma reesei* proteome with prediction score 1.

Sequence	Hydropathy	Net Charge	Helix Fraction	Turn Fraction	Sheet Fraction	Instability index
KQKKKRSRALHRQRVDNRRS	-2.58	8.58435271	0.1	0.15	0.1	121.945
QKKKRSALHRQRVDNRRSS	-2.425	7.58685831	0.1	0.2	0.1	119.755
RRRNTAASARFRIKKKQREQ	-2.075	7.55090614	0.1	0.1	0.2	85.91
KRRKTGAVPSLTRRVYLRLR	-0.935	7.54978538	0.3	0.15	0.2	109.34
KKRYKAYGKWLRGKSFFKQQ	-1.665	7.53735789	0.3	0.15	0.1	34.785
RKAGRSCVVTAPTRKRQKKT	-1.375	7.52282666	0.1	0.15	0.1	75.315
PSETGGIRKRKRKEKKRQWV	-2.205	6.89094115	0.15	0.2	0.1	80.725
AKRKRPKWKIVPLEISSRGR	-1.22	6.59880022	0.25	0.25	0.15	77.515
HKRSASGAKRAYYRKKRAFE	-1.745	6.58215853	0.15	0.15	0.25	85.68
KGRVASHPKKTTPSKKPKAS	-1.615	6.57902063	0.05	0.35	0.1	9.675

Table 19. BlastP results for the first 10 peptides with the highest net charge (20 aminoacid long peptides) predicted from *Trichoderma reesei* proteome with prediction score 1.

Sequence	First hit of BlastP	Secondary and other hits
KQKKKRSRALHRQRVDNRRS	uncharacterized protein TRIREDRAFT_103812 [*Trichoderma reesei* QM6a]	Transcription initiation factor IIB [*Trichoderma lentiforme*]
QKKKRSRALHRQRVDNRRSS	uncharacterized protein TRIREDRAFT_103812 [*Trichoderma reesei* QM6a]	Transcription initiation factor IIB [*Trichoderma lentiforme*]
RRRNTAASARFRIKKKQREQ	hypothetical protein EJ08DRAFT_563690 [*Tothia fuscella*]	BQ2448_2821 [*Microbotryum intermedium*]
KRRKTGAVPSLTRRVYLRLR	uncharacterized protein TRIREDRAFT_104334 [*Trichoderma reesei* QM6a]	nitrogenase [*Erysipelotrichaceae bacterium*]
KKRYKAYGKWLRGKSFFKQQ	polyphosphoinositide phosphatase, putative [*Trichoderma parareesei*]	
RKAGRSCVVTAPTRKRQKKT	hypothetical protein M440DRAFT_1344644 [*Trichoderma longibrachiatum* ATCC 18648]	C6 transcription factor (War1) [*Purpureocillium lilacinum]*
PSETGGIRKRKRKEKKRQWV	hypothetical protein BBK36DRAFT_1158309 [*Trichoderma citrinoviride*]	Select seq ref|XP_036576375.1| glucan 4-alpha-glucosidase [*Colletotrichum truncatum*]
AKRKRPKWKIVPLEISSRGR	uncharacterized protein TRIREDRAFT_102933 [*Trichoderma reesei* QM6a]	30S ribosomal protein S17 [*Elusimicrobia bacterium*]
HKRSASGAKRAYYRKKRAFE	hypothetical protein AC578_1324 [*Pseudocercospora eumusae]*	40S ribosomal protein S8-B [*Pseudocercospora fuligena*]
KGRVASHPKKTTPSKKPKAS	d-lactate dehydrogenase 2, mitochondrial precursor [*Trichoderma arundinaceum*]	

Table 20. First 10 peptides (20 aminoacid long peptides) with the highest hydropathy and their physicochemical, structural properties predicted from *Trichoderma reesei* proteome with prediction score 1.

Sequence	Hydropathy	Net Charge	Helix Fraction	Turn Fraction	Sheet Fraction	Instability index
NLVCAALLFISIVMGIFLLE	2.315	-1.4638373	0.6	0.15	0.45	1.01
PRWVIVACLLMVIVGLLALY	2.205	0.87429398	0.65	0.1	0.4	48.99
TLIIVEVIPFIVLAVGVDNI	2.125	-2.7900404	0.65	0.15	0.2	19.41
VCAALLFISIVMGIFLLEET	2.09	-2.499111	0.55	0.1	0.45	17.31
LPILLCAAVAFLVIPDDIGL	2.05	-2.4661575	0.55	0.15	0.4	23.55
RPRWVIVACLLMVIVGLLAL	2.045	1.53289926	0.6	0.1	0.4	58.62
ILPLFTKAAFFWSVAGFVII	1.89	0.55494727	0.6	0.15	0.25	28.76
NAAKAVAGVLPAAVLALLLF	1.88	0.55494727	0.45	0.15	0.6	28.76
ASLHWVALLFMVLSLGVFFN	1.795	-0.3555713	0.6	0.2	0.4	3.815
LHILSVAFDACACMLFSALG	1.745	-1.4544809	0.4	0.15	0.45	107.51

Table 21. BlastP results for the first 10 peptides with the highest hydropathy (20 aminoacid long peptides) predicted from *Trichoderma reesei* proteome with prediction score 1.

Sequence	First hit of BlastP	Secondary and other hits
NLVCAALLFISIVMGIFLLE	MFS permease [*Trichoderma parareesei*]	
PRWVIVACLLMVIVGLLALY	hypothetical protein M419DRAFT_125607 [*Trichoderma reesei* RUT C-30]	tetratricopeptide repeat protein [*Mucilaginibacter* sp. E4BP6]
TLIIVEVIPFIVLAVGVDNI	MFS permease [*Trichoderma parareesei*]	
VCAALLFISIVMGIFLLEET	MFS permease [*Trichoderma parareesei*]	
LPILLCAAVAFLVIPDDIGL	MFS general substrate transporter [*Trichoderma reesei* RUT C-30]	
RPRWVIVACLLMVIVGLLAL	hypothetical protein M419DRAFT_125607 [*Trichoderma reesei* RUT C-30]	glycosyl transferase [Burkholderiaceae bacterium]
ILPLFTKAAFFWSVAGFVII	gaba permease [*Trichoderma reesei* RUT C-30]	
NAAKAVAGVLPAAVLALLLF	DUF221-domain-containing protein [*Trichoderma longibrachiatum* ATCC 18648]	
ASLHWVALLFMVLSLGVFFN	Zn2Cys6 transcriptional regulator [*Trichoderma parareesei*]	
LHILSVAFDACACMLFSALG	NRPS protein [*Trichoderma parareesei*]	acetyl-CoA synthetase-like protein [*Trichoderma reesei* RUT C-30]

Table 22. First 10 peptides (30 aminoacid long peptides) and their physicochemical, structural properties predicted from *Trichoderma reesei* proteome with prediction score 1.

Sequence	Hydropathy	Net Charge	Helix Fraction	Turn Fraction	Sheet Fraction	Instability index
SYPRHLILAIIAKLPSYIKLSIVKRMLAHI	0.74	4.31390299	0.46666667	0.16666667	0.3	47.35
QGRAGRVQAGKCFKLFTQNLEQNMAPRPEP	-0.96	2.53010448	0.16666667	0.26666667	0.26666667	29.5566667
GRAGRVQAGKCFKLFTQNLEQNMAPRPEPE	-0.96	1.53331773	0.16666667	0.26666667	0.3	35.35
SYPRHLILAIIAKLPSYIKLSIVKRMLAHI	0.74	4.31390299	0.46666667	0.16666667	0.3	47.35
QGRAGRVQAGKCFKLFTQNLEQNMAPRPEP	-0.96	2.53010448	0.16666667	0.26666667	0.26666667	29.5566667
GRAGRVQAGKCFKLFTQNLEQNMAPRPEPE	-0.96	1.53331773	0.16666667	0.26666667	0.3	35.35
FALVWFVATEFLRRTGWLAWGLELPPAKML	0.69	0.55713856	0.46666667	0.13333333	0.43333333	73.6666667
CYCWRKDMNPVVYAHANNTCVATVALARLI	0.35666667	1.51344868	0.33333333	0.13333333	0.26666667	20.2266667
SAPQEPVWAKSWITGFYHQMRALLYRPSPQ	-0.6	1.2834085	0.3	0.26666667	0.23333333	74.6636667
ALAPASESEFNEVIKAELLDPNATAILDNI	0.18666668	-5.3893273	0.3	0.23333333	0.46666667	60.0033333

Table 23. BlastP results for the first 10 peptides (30 aminoacid long peptides) predicted from *Trichoderma reesei* proteome with prediction score 1.

Sequence	First hit of BlastP	Secondary and other hits
SYPRHLILAIIAKLPSYIKLSIVKRMLAHI	Uncharacterized protein TRIREDRAFT_53044 [*Trichoderma reesei* QM6a]	P-loop containing nucleoside triphosphate hydrolase protein [*Trichoderma citrinoviride*]
QGRAGRVQAGKCFKLFTQNLEQNMAPRPEP	uncharacterized protein TRIREDRAFT_53044 [*Trichoderma reesei* QM6a]	P-loop containing nucleoside triphosphate hydrolase protein [*T. longibrachiatum*]
GRAGRVQAGKCFKLFTQNLEQNMAPRPEPE	uncharacterized protein TRIREDRAFT_53044 [*Trichoderma reesei* QM6a]	P-loop containing nucleoside triphosphate hydrolase protein [*T. longibrachiatum*
SYPRHLILAIIAKLPSYIKLSIVKRMLAHI	uncharacterized protein TRIREDRAFT_53044 [*Trichoderma reesei* QM6a]	P-loop containing nucleoside triphosphate hydrolase protein [*T. citrinoviride*]
QGRAGRVQAGKCFKLFTQNLEQNMAPRPEP	uncharacterized protein TRIREDRAFT_53044 [*Trichoderma reesei* QM6a]	putative ATP-dependent RNA helicase ucp12 [*Trichoderma lentiforme*]
GRAGRVQAGKCFKLFTQNLEQNMAPRPEPE	uncharacterized protein TRIREDRAFT_53044 [*T. reesei* QM6a]	P-loop containing nucleoside triphosphate hydrolase protein [*T. longibrachiatum*]
FALVWFVATEFLRRTGWLAWGLELPPAKML	uncharacterized protein TRIREDRAFT_42193 [*T. reesei* QM6a]	PAP2-domain-containing protein [T. *longibrachiatum* ATCC 18648]
CYCWRKDMNPVVYAHANNTCVATVALARLI	uncharacterized protein TRIREDRAFT_53499 [*T. reesei* QM6a]	UvrD-helicase domain-containing protein [*Scopulibacillus daqui*]
SAPQEPVWAKSWITGFYHQMRALLYRPSPQ	Zn2Cys6 transcriptional regulator [*T. parareesei]*	
ALAPASESEFNEVIKAELLDPNATAILDNI	uncharacterized protein TRIREDRAFT_73560 [*T. reesei* QM6a]	UTP20, snoRNA-binding protein and member of the SSU processome [*T. parareesei*]

Table 24. First 10 peptides (30 aminoacid long peptides) with the highest net charge and their physicochemical, structural properties predicted from *Trichoderma reesei* proteome with prediction score 1.

Sequence	Hydropathy	Net Charge	Helix Fraction	Turn Fraction	Sheet Fraction	Instability index
LRTITRHHKQIRRMYERAKNTNSAFFWGKR	-1.3566667	7.62165343	0.23333333	0.13333333	0.16666667	59.4933333
RCGKGRSYTSWARQDVHESKRRKTTSQKTG	-1.8733333	6.55848963	0.1	0.23333333	0.06666667	63.6233333
PIYKTEPWNQWVKAILNMPIKRDVKVRRGR	-0.8833333	5.89026687	0.33333333	0.2	0.13333333	60.9333333
RVVHHHQQRKPTLKDKVTGALTRLKGSLTG	-0.8666667	5.65768095	0.23333333	0.16666667	0.16666667	-11.896667
RCRAVRLAIKCMQHLEGFGRNAHNRVKKDP	-0.83	5.57562389	0.2	0.16666667	0.23333333	23.4066667
AQKCDTLAGWWTLLLEKEERKQRRRERKRK	-1.7766667	5.57554979	0.2	0.03333333	0.33333333	77.98
GHSCYRPRRTGERKRKSVRGCIVAMDLSVL	-0.6266667	5.53897233	0.23333333	0.23333333	0.16666667	37.76
RNCRRALKWQFVIGQSVLRLLPLAYFWLKP	0.12	5.52533225	0.46666667	0.16666667	0.26666667	52.8266667
MRNCRRALKWQFVIGQSVLRLLPLAYFWLK	0.23666667	5.25276786	0.46666667	0.13333333	0.3	52.8266667
YAKHLSWNNWASIGPPTRIVPFRLLQRLFK	-0.2666667	4.58648718	0.4	0.26666667	0.2	35.4933333

Table 25. BlastP results of first 10 peptides (30 aminoacid long peptides) with the highest net charge predicted from *Trichoderma reesei* proteome with prediction score 1.

Sequence	First hit of BlastP	Secondary and other hits
LRTITRHHKQIRRMYERAKNTNSAFFWGKR	uncharacterized protein TRIREDRAFT_104372 [*Trichoderma reesei* QM6a]	ABC transporter permease [*Faecalicatena fissicatena]*
RCGKGRSYTSWARQDVHESKRRKTTSQKTG	uncharacterized protein TRIREDRAFT_103898 [*Trichoderma reesei* QM6a]	helix-turn-helix domain-containing protein [*Occultella aeris*]
PIYKTEPWNQWVKAILNMPIKRDVKVRRGR	uncharacterized protein TRIREDRAFT_103335 [*Trichoderma reesei* QM6a]	
RVVHHHQQRKPTLKDKVTGALTRLKGSLTG	uncharacterized protein TRIREDRAFT_103335 [*Trichoderma reesei* QM6a]	
RCRAVRLAIKCMQHLEGFGRNAHNRVKKDP	6-phosphofructokinase-like protein [Trichoderma citrinoviride]	
AQKCDTLAGWWTLLLEKEERKQRRRERKRK	hypothetical protein M419DRAFT_6474 [*Trichoderma reesei* RUT C-30]	kinase-like protein [*Trichoderma citrinoviride*]
GHSCYRPRRTGERKRKSVRGCIVAMDLSVL	6-phosphofructokinase-like protein [*Trichoderma citrinoviride*]	
RNCRRALKWQFVIGQSVLRLLPLAYFWLKP	TUL1 Golgi-localized RING-finger ubiquitin ligase (E3) [*Trichoderma parareesei*]	trans-membrane Golgi-localized RING-finger ubiquitin ligase [*Trichoderma reesei* QM6a]
MRNCRRALKWQFVIGQSVLRLLPLAYFWLK	TUL1 Golgi-localized RING-finger ubiquitin ligase (E3) [*Trichoderma parareesei*]	trans-membrane Golgi-localized RING-finger ubiquitin ligase [*Trichoderma reesei* QM6a]
YAKHLSWNNWASIGPPTRIVPFRLLQRLFK	uncharacterized protein TRIREDRAFT_120879 [*Trichoderma reesei* QM6a]	alpha/beta-hydrolase [*Trichoderma longibrachiatum* ATCC 18648]

Table 26. First 10 peptides (30 aminoacid long peptides) with the highest hydropathy and their physicochemical, structural properties predicted from *Trichoderma reesei* proteome with prediction score 1.

Sequence	Hydropathy	Net Charge	Helix Fraction	Turn Fraction	Sheet Fraction	Instability index
ICNFAAYAFAPAILVTPLGALSVLIGAVLG	1.83666667	-0.4695561	0.46666667	0.23333333	0.4	23.32
VLRCLVGPFMFFAFAVLVKLLLDAVLGKMT	1.72	1.49405763	0.53333333	0.1	0.4	29.3833333
SEVFGRALIRPWVLLFREPIVLIASLYMAI	1.15333333	0.25288889	0.53333333	0.16666667	0.36666667	41.0666667
GFGALLGFASACIVWSEIVKFPFPKLSIIR	1.11333333	1.52903396	0.46666667	0.26666667	0.23333333	44.1333333
YCTFTAYLAFNFLCAFAPNFGSLLVGRFLT	1.09333333	0.50340983	0.46666667	0.2	0.3	18.2166667
DLDWAYNMRLLLLSFLISGIFTISNYFLYPI	0.98666667	-1.4466905	0.56666667	0.23333333	0.3	5.73333333
TWCSRCATGLSMLVLIGLCYWVISRESMVE	0.84666667	-0.8633395	0.4	0.2	0.3	14.8
LGLHVTQIRILVRALQLLKVGGRVVYSTCS	0.8	3.5644894	0.46666667	0.16666667	0.23333333	14.72
SYPRHLILAIIAKLPSYIKLSIVKRMLAHI	0.74	4.31390299	0.46666667	0.16666667	0.3	47.35
KLKVFAQWIANAIYHSLLLYVFAELIWYGD	0.72666667	-0.4156121	0.56666667	0.1	0.33333333	2.46

Table 27. BlastP results of first 10 peptides (30 aminoacid long peptides) with the highest hydropathy and their physicochemical, structural properties predicted from *Trichoderma reesei* proteome with prediction score 1.

Sequence	First hit of BlastP	Secondary and other hits
ICNFAAYAFAPAILVTPLGALSVLIGAVLG	uncharacterized protein CkaCkLH20_10868 [*Colletotrichum karsti*]	DUF803 domain membrane protein [*Akanthomyces lecanii* RCEF 1005]
VLRCLVGPFMFFAFAVLVKLLLDAVLGKMT	uncharacterized protein TRIREDRAFT_55306 [*Trichoderma reesei* QM6a]	peroxisomal-coenzyme A synthetase [*Trichoderma harzianum]*
SEVFGRALIRPWVLLFREPIVLIASLYMAI	MFS multidrug transporter [*Trichoderma parareesei*]	
GFGALLGFASACIVWSEIVKFPFPKLSIIR	LMBR1 domain-containing protein [*Trichoderma parareesei*]	integral membrane [*Fusarium agapanthi*]
YCTFTAYLAFNFLCAFAPNFGSLLVGRFLT	uncharacterized protein TRIREDRAFT_21595 [*Trichoderma reesei* QM6a]	MFS general substrate transporter [*Trichoderma reesei* RUT C-30]
DLDWAYNMRLLLLSFLISGIFTISNYFLPI	OPT superfamily oligopeptide transporter [*Trichoderma reesei* RUT C-30]	
TWCSRCATGLSMLVLIGLCYWVISRESMVE	hypothetical protein A0O28_0068960 [*Trichoderma guizhouense]*	glycosyl transferase GT-A type structural fold protein [*Metarhizium robertsii*]
LGLHVTQIRILVRALQLLKVGGRVVYSTCS	hypothetical protein CFAM422_005207 [*Trichoderma lentiforme*]	methyltransferase [*Trichoderma arundinaceum*]
SYPRHLILAIIAKLPSYIKLSIVKRMLAHI	uncharacterized protein TRIREDRAFT_53044 [*Trichoderma reesei* QM6a]	DEAD/DEAH box helicase [*Metarhizium brunneum* ARSEF 3297]
KLKVFAQWIANAIYHSLLLYVFAELIWYGD	uncharacterized protein TRIREDRAFT_43831 [*Trichoderma reesei* QM6a]	phospholipid-translocating P-type ATPase [*Trichoderma parareesei*]

As it can be observed from the predicted AMPs and their corresponding BlastP hits, CNN-BiLSTM model can predict a diverse set of AMPs which comes from a rich set of microbial organisms such as bacteria, protezoa, fungi. Overall, as can be seen from the results of deep learning models, our model is learning more generally. The model makes predictions for hydrophobic transmembrane protein regions and peptide fragments rich in positive charge. Structure predictions were applied to someof the representative peptides with PEPFOLD3 (Lamiable et al., 2016). The secondary structure predictions show that these peptides are mostly alpha helical with a few of the selected have minimal beta sheet structures (Figure 18, Figure 19, Figure 20, Figure 21, Figure 22). All of this shows that our deep learning methodology can learn well, despite being trained on a heterogeneous and more challenging dataset than the literature used.

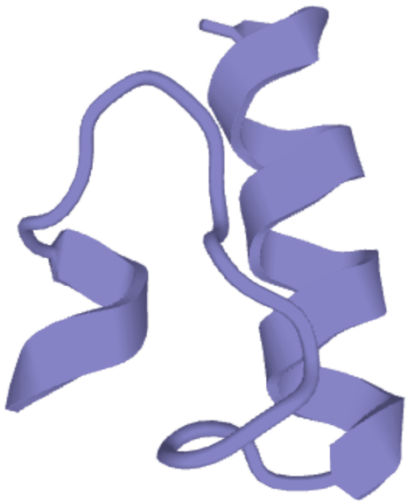


Figure 18. PepFOLD3 prediction for SYPRHLILAIIAKLPSYIKLSIVKRMLAHI

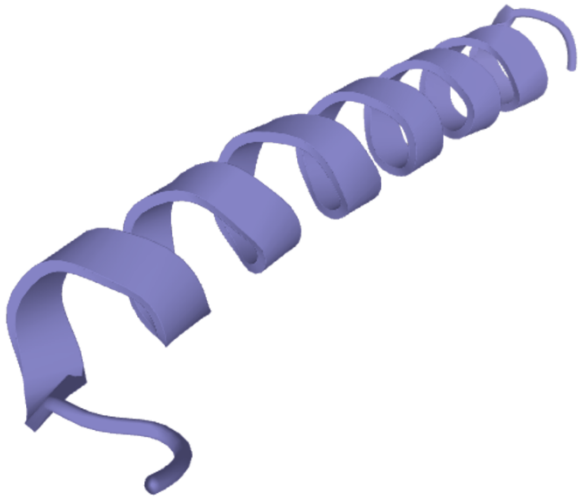


Figure 19. PepFOLD3 prediction for IVGKKYAARIQELSLAVYSAAAAYALERGI

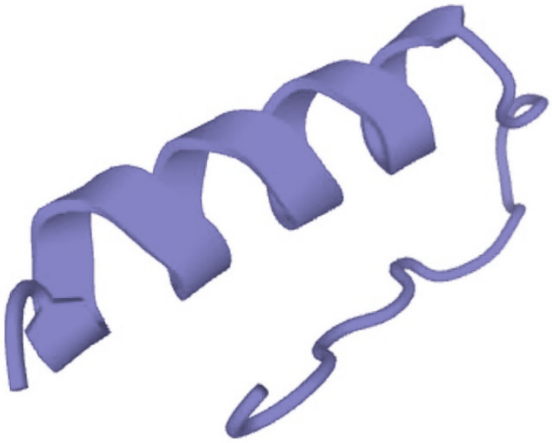


Figure 20. PepFOLD3 prediction for LGLHVTQIRILVRALQLLKVGGRVVYSTCS

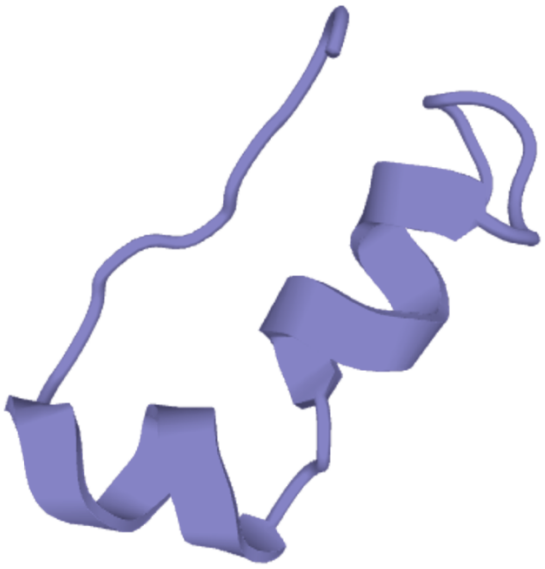


Figure 21. PepFOLD3 prediction for FALVWFVATEFLRRTGWLAWGLELPPAKML

Figure 22. PepFOLD3 prediction for RCGALARVATSSTKLLSARAAFPASFSGAA

A diverse set of peptides were selected from the prediction pool by evaluating the presence of high positive charge, hydropathy, availability of Cysteine residues, low instability index, secondary structure and non-toxicity against human. Table 28 enlists the diversely selected peptides and their amino acid sequences. Various physico-chemical parameters and above-mentioned parameters were shown in Table 29. Selected peptides were blasted using BlastP. A functional analysis of the hits was employed to select for the peptides with promising AMP properties. The similarity of these peptides against the human proteome was also evaluated in order to minimize interference with the human systems for future use. Table 30 exhibits the first four BLASTP hits against the non-redundant protein database and the related search score results. Table 31 shows the hits against the human proteome. As seen from aforementioned tables, the similarities against peptides from microbial sources are high whereas the similarities against human proteome are low. Using these criteria in the selection of the peptides with AMP activity for use in human systems could minimize their potential off-target effects.

PEP-FOLD3 was used in the modelling of the selected peptides. Table 32 indicates PepFold-3 cluster modelling sizes and secondary structure probability graph for the selected peptides. Modelled peptides were represented in Figure 23. The selected AMP set, overwhelmingly exhibit alpha helical structure whereas one AMP shows minimal beta-sheet structure. These structures corroborate the known AMP structures in the available literature.

Table 28. Diverse peptides and their sequences selected for experimentation.

Internal Identifier(ID)	Sequence
CYS-AMP1	MKPCALICASKAARICRQIL
CYS-AMP2	CATCIASSQCQPTRYRLSFG
CYS-AMP3	RCSEIHCFCFFIAGVYLHVL
CHARGE-AMP1	VAKKAVKKKAVKKKAVKKKV
CHARGE-AMP2	PPRARFRAARIARARGLRTT
CHARGE-AMP3	RAVLVHRVRILTKKDRKRLY
HYDROPHOBIC-AMP1	MKALVILGFLFLSVAVQGKV
HYDROPHOBIC-AMP2	MGTLVKAAQAFLFSLVLILS

Table 29. Properties of the selected peptides.

ID	pI	Hydropathy	Charge	Cysteine count	Helix Fraction	Turn Fraction	Sheet Fraction	Instability Index	Toxicity Prediction *
CYS-AMP1	9,61	0,75	3,21	3,00	0,25	0,10	0,35	21,22	Non-Toxin
CYS-AMP2	8,71	-0,05	1,48	3,00	0,20	0,25	0,15	18,63	Non-Toxin
CYS-AMP3	6,89	1,29	-0,44	3,00	0,50	0,10	0,20	31,16	Non-Toxin
CHARGE-AMP1	11	-0,74	10,50	0	0,25	0,00	0,20	0,94	Non-Toxin
CHARGE-AMP2	12,78	-0,82	6,90	0	0,15	0,15	0,30	20,75	Non-Toxin
CHARGE-AMP3	11,84	-0,63	6,58	0	0,40	0,00	0,20	31,66	Non-Toxin
HYDROPHOBIC-AMP1	10	1,74	1,28	0	0,55	0,15	0,35	10,64	Non-Toxin
HYDROPHOBIC-AMP2	8,5	1,74	0,28	0	0,50	0,15	0,45	13,45	Non-Toxin

Table 30. BlastP results of the selected peptides against the non-redundant protein database.

ID	BlastP Hit	Maximum Score	Total Score	Query Coverage	E-value	% of Identity	Accession Length	Accession Identifier
CYS-AMP1	hypothetical protein CSQ95_20895 [*Janthinobacterium* sp. BJB304]	68.9	68.9	100%	7e-12	100.00%	95	PHV37208.1
	hypothetical protein [*Janthinobacterium* sp. HH106]	68.9	68.9	100%	7e-12	100.00%	113	WP_139142888.1
	penicillin-binding protein 1C [*Maridesulfovibrio bastinii*]	33.7	33.7	100%	17	55.00%	782	WP_027180309.1
	FtsX-like permease family protein [*Spirochaetes* bacterium]	32.9	32.9	60%	33	91.67%	420	MBU1076834.1
CYS-AMP2	hypothetical protein CLF_106301 [*Clonorchis sinensis*]	68.5	68.5	100%	9e-12	100.00%	1338	GAA54983.1
	hypothetical protein SAMD00023353_0801770 [*Rosellinia necatrix*]	35.0	35.0	85%	5.9	64.71%	597	GAW25539.1
	hypothetical protein [C*orynebacteriales* bacterium]	34.1	34.1	85%	11	63.16%	98	NUS72212.1
	hypothetical protein [*Paramecium tetraurelia* strain d4-2]	34.1	82.9	75%	12	64.71%	1650	XP_001431988.1
CYS-AMP3	PREDICTED: zinc transporter 5-like [*Notothenia coriiceps*]	72.3	72.3	100%	4e-13	100.00%	188	XP_010778531.1
	HAD-IB family phosphatase [Cl*ostridium* sp. C1]	35.8	35.8	75%	2.9	65.00%	200	WP_211964867.1
	serine/threonine protein phosphatase [*Ruminococcus* sp.]	35.0	35.0	85%	5.9	53.85%	225	MBQ9894850.1
	serine/threonine protein phosphatase [*Ruminococcus* sp.]	35.0	35.0	85%	5.9	53.85%	225	MBR6394507.1
CHARGE-AMP1	histidine biosynthesis protein HisIE [*Puniceicoccaceae bacterium*]	60.9	212	100%	2e-09	100.00%	46	MAW20641.1
	histidine biosynthesis protein HisIE [*Puniceicoccaceae bacterium*]	60.9	246	100%	3e-09	100.00%	60	MBH78754.1
	hypothetical protein [*Candidatus Woesearchaeota* archaeon]	57.9	181	100%	5e-08	95.00%	75	MBT4824611.1
	histone [*Candidatus Woesearchaeota* archaeon]	55.4	779	100%	4e-07	100.00%	132	MBT7903001.1

Table 30. BlastP results of the selected peptides against the non-redundant protein database (Continued).

ID	BlastP Hit	Maximum Score	Total Score	Query Coverage	E-value	% of Identity	Accession Length	Accession Identifier
CHARGE-AMP2	membrane protein [*Burkholderia latens*]	65.1	65.1	100%	1e-10	100.00%	319	VWB49790.1
	cytochrome c oxidase assembly protein [*Burkholderia latens*]	65.1	65.1	100%	1e-10	100.00%	294	WP_174904761.1
	cytochrome c oxidase assembly protein [*Burkholderia latens*]	65.1	65.1	100%	1e-10	100.00%	294	WP_175923050.1
	cytochrome c oxidase assembly protein [*Burkholderia latens*]	65.1	65.1	100%	1e-10	100.00%	160	WP_151068810.
CHARGE-AMP3	hypothetical protein [*Armatimonadetes* bacterium]	34.6	34.6	60%	7.9	83.33%	59	MBN9500516.1
	NERD domain-containing protein [*Bacillus* sp. BGMRC 2118]	34.6	34.6	50%	8.3	100.00%	322	KAA0548359.1
	hypothetical protein A3F54_04345 [*Candidatus Kerfeldbacteria* bacterium RIFCSPHIGHO2_12_FULL_48_17]	34.6	34.6	65%	8.3	84.62%	515	OGY85128.1
	polysaccharide deacetylase family protein [*Aeriscardovia* sp.]	34.1	34.1	95%	12	70.00%	250	MBR2555554.1
HYDROPHOBIC-AMP1	PREDICTED: lysozyme C-2-like [*Bos indicus*]	64.7	64.7	100%	2e-10	100.00%	167	XP_019816395.1
	RecName: Full=Lysozyme C; AltName: Full=1,4-beta-N-acetylmuramidase C; Flags: Precursor [*Bos taurus*]	64.7	64.7	100%	2e-10	100.00%	147	P04421.2
	lysozyme C-2 precursor [*Bos taurus]*	64.7	64.7	100%	2e-10	100.00%	147	NP_851342.1
	TPA: lysozyme F5 [*Bos taurus*]	62.6	62.6	100%	1e-09	95.00%	167	CDM98814.1
HYDROPHOBIC-AMP2	hypothetical protein AUF67_04745 [Acidobacteria bacterium 13_1_20CM_58_21]	64.3	64.3	100%	3e-10	100.00%	503	OLD82294.1
	hypothetical protein AUG46_04995 [Acidobacteria bacterium 13_1_20CM_3_58_11]	64.3	64.3	100%	3e-10	100.00%	380	OLE47908.1

Table 30. BlastP results of the selected peptides against the non-redundant protein database (Continued).

ID	BlastP Hit	Maximum Score	Total Score	Query Coverage	E-value	% of Identity	Accession Length	Accession Identifier
	hypothetical protein DMG44_08925 [Acidobacteria bacterium]	43.9	43.9	65%	0.004	100.00%	503	PYT49947.1
	transporter substrate-binding domain-containing protein [*Loigolactobacillus backii*]	36.7	36.7	55%	1.5	100.00%	275	WP_068222710.1

Table 31. BlastP results of the selected peptides against the human proteome.

ID	Description	Maximum Score	Total Score	Query Coverage	E-value	% of Identity	Accession Length	Accession Number
CYS-AMP1	ALKBH1 protein [*Homo sapiens*]	26.1	26.1	90%	8.2	44.44%	380	AAH15024.1
	Unknown [*Homo sapiens*]	23.5	23.5	35%	37	85.71%	18	AAD13933.1
	immunoglobulin light chain junction region [*Homo sapiens*]	22.7	22.7	40%	52	87.50%	12	MBB1656375.1
	unknown [Homo sapiens]	23.5	23.5	50%	64	80.00%	198	AAY15035.1
CYS-AMP2	immunoglobulin heavy chain junction region [*Homo sapiens*]	25.2	25.2	60%	7.1	75.00%	15	MOL60324.1
	immunoglobulin heavy chain junction region [*Homo sapiens*]	25.7	25.7	80%	7.6	62.50%	23	MBN4558951.1
	immunoglobulin heavy chain junction region [*Homo sapiens*]	24.8	24.8	90%	16	50.00%	23	MCG23393.1
	immunoglobulin heavy chain junction region [*Homo sapiens*]	24.4	24.4	50%	18	70.00%	18	MBK4192006.1
CYS-AMP3	unnamed protein product [*Homo sapiens*]	26.1	26.1	45%	8.2	58.33%	374	BAA92103.1
	F-box/LRR-repeat protein 8 [*Homo sapiens*]	26.1	26.1	45%	8.2	58.33%	374	NP_060848.2
	immunoglobulin heavy chain junction region [*Homo sapiens*]	24.8	24.8	50%	12	70.00%	18	MOK19791.1
	laminin, alpha 1, isoform CRA_c [*Homo sapiens*]	25.2	25.2	70%	17	64.29%	2506	EAX01630.1
CHARGE-AMP1	MAP1B protein [*Homo sapiens*]	29.9	164	100%	0.37	50.00%	727	AAH94834.1
	histone H1.5 [*Homo sapiens*]	29.1	187	95%	0.73	65.22%	226	NP_005313.1
	CYLC1 protein [*Homo sapiens*]	28.6	98.6	90%	1.0	61.90%	334	AAH58019.1
	histone H1.2 [*Homo sapiens*]	28.2	149	100%	1.5	62.50%	213	NP_005310.1

Table 31. BlastP results of the selected peptides against the human proteome (Continued).

ID	Description	Maximum Score	Total Score	Query Coverage	E-value	% of Identity	Accession Length	Accession Number
CHARGE-AMP2	zinc finger protein 706, isoform CRA_b [*Homo sapiens*]	29.1	29.1	75%	0.73	73.33%	162	EAW91832.1
	hCG28655 [*Homo sapiens*]	26.9	26.9	80%	4.2	65.00%	556	EAX02595.1
	immunoglobulin heavy chain junction region [*Homo sapiens*]	26.1	26.1	45%	5.5	88.89%	24	MOQ99574.1
	Structure of the human PKD1/PKD2 complex [*Homo sapiens*]	26.5	41.1	85%	5.9	64.29%	1153	6A70_B
CHARGE-AMP3	leucine zipper transcription factor-like protein 1 isoform 2 [*Homo sapiens*]	25.7	25.7	50%	12	56.25%	282	NP_001263307.1
	Leucine zipper transcription factor-like 1 [*Homo sapiens*]	25.7	25.7	50%	12	56.25%	299	AAH25988.1
	leucine zipper transcription factor-like protein 1 isoform 1 [*Homo sapiens*]	25.7	25.7	50%	12	56.25%	299	NP_065080.1
	leucine zipper transcription factor-like 1 [*Homo sapiens*]	25.7	25.7	50%	12	56.25%	299	CAB96873.1
HYDROPHOBIC-AMP1	lysozyme C precursor [*Homo sapiens*]	43.1	43.1	100%	9e-06	70.00%	148	NP_000230.1
	lysozyme precursor (EC 3.2.1.17) [*Homo sapiens]*	43.1	43.1	100%	9e-06	70.00%	148	AAA36188.1
	lysozyme precursor [*Homo sapiens*]	41.4	41.4	100%	3e-05	70.00%	148	AAC63078.1
	protocadherin beta 1 [*Homo sapiens*]	30.3	46.7	65%	0.26	78.57%	818	AAD43749.1
HYDRO PHOBI	transmembrane protein 87B isoform X2 [*Homo sapiens*]	26.5	26.5	40%	5.8	87.50%	354	XP_016860610.1

Table 31. BlastP results of the selected peptides against the human proteome (Continued).

ID	Description	Maximum Score	Total Score	Query Coverage	E-value	% of Identity	Accession Length	Accession Number
	unknown [*Homo sapiens*]	26.5	26.5	40%	5.9	87.50%	526	AAY24214.1
	transmembrane protein 87B isoform X1 [*Homo sapiens*]	26.5	26.5	40%	5.9	87.50%	554	XP_005263884.1
	transmembrane protein 87B isoform 1 precursor [*Homo sapiens*]	26.5	26.5	40%	5.9	87.50%	555	NP_116213.1

Overall Blast results demonstrates a wide range of plausible mechanisms with which these candidate AMPs can act antimicrobial towards pathogens. Most of the Blast results gave hit to histon proteins, transcription binding factors, membrane transports and channel proteins and to parts of signal peptides. It is known that the positively charged histon proteins and transcription binding factors have high affinity to bind to DNA regions in order to kill bacterias and fungus by disabling vital transcription mechanisms in the DNA. Human histones and antimicrobial proteins are found to be in close vicinity in microbial pathogenesis and this points to a related mechanism. It is used to be postulated that histones acted as AMPs before their primary functions are researched. Histone proteins are known to further disintegrates the pores opened by AMPs, prevent bacterial membrane regeneration, depolarize bacterial proton gradient, reconstructing bacterial chromosomal arrangement by entering into cytoplasm and prevent transcription.

Similarly, hydrophobic regions of membrane transporter and channel protein fragments are shown to have antibacterial and antifungal properties. As signal peptides has high hydropathy scores and that they can embed alpha helical secondary structures into the membrane during protein secretion. In current literature, there are many examples of mutated versions of signal peptides which have strong antimicrobial properties. In similar manner, positively charged linear cationic AMPs that are similar to the positively charged peptides predicted by the model are well known and well

researched. The most typical example is LL-37 which shows antimicrobial activity through disruption transcriptional regulation.

Table 32. PepFold-3 cluster modelling sizes and secondary structure probability graph for the selected peptides.

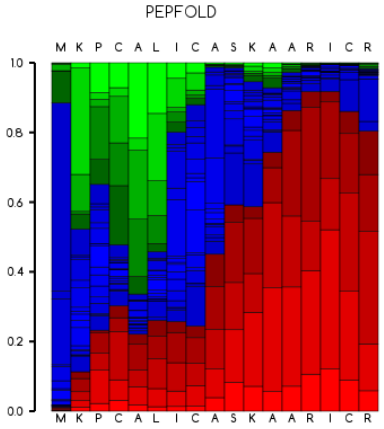
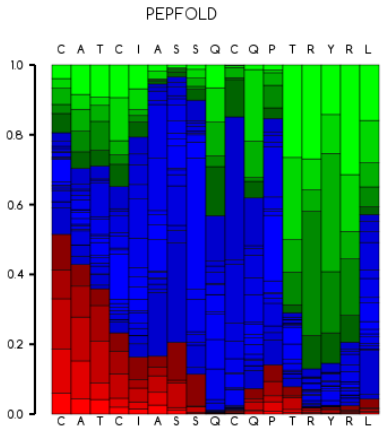
ID	Size of Cluster 1	Total # of Clusters	Pep-Fold3 Secondary Structure Probability Graph (Red: Helix, Green: Extended, Blue: Random Coil)
CYS-AMP1	62	9	
CYS-AMP2	1	53	

Table 32. PepFold-3 cluster modelling sizes and secondary structure probability graph for the selected peptides (Continued).

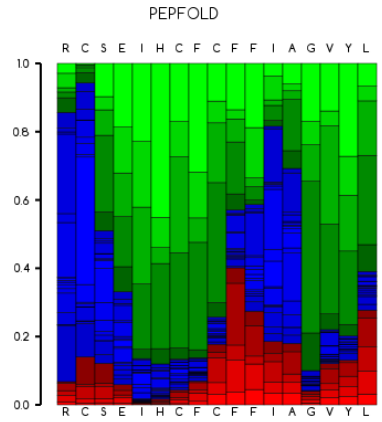
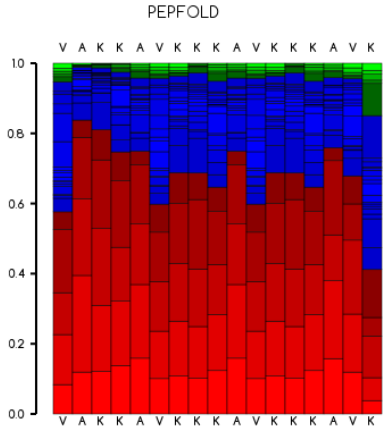
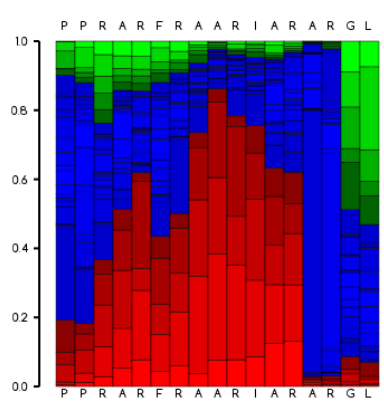
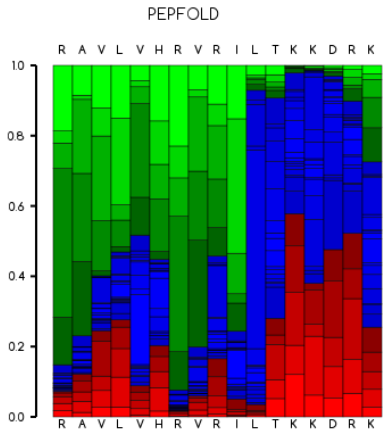
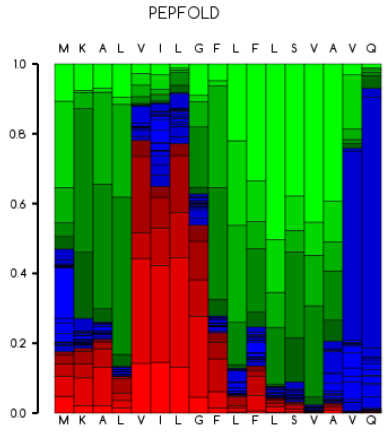
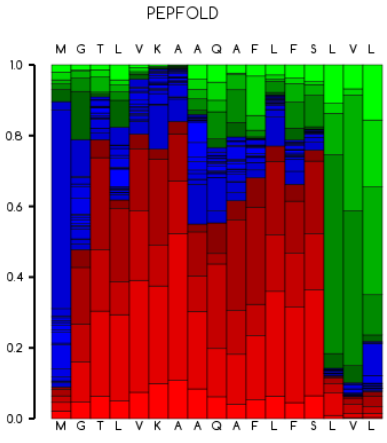
ID	Size of Cluster 1	Total # of Clusters	Pep-Fold3 Secondary Structure Probability Graph (Red: Helix, Green: Extended, Blue: Random Coil)
CYS-AMP3	4	74	
CHARGE-AMP1	23	10	
CHARGE-AMP2	19	7	

Table 32. PepFold-3 cluster modelling sizes and secondary structure probability graph for the selected peptides (Continued).

ID	Size of Cluster 1	Total # of Clusters	Pep-Fold3 Secondary Structure Probability Graph (Red: Helix, Green: Extended, Blue: Random Coil)
CHARGE-AMP3	2	44	
HYDROPHOBIC-AMP1	2	71	
HYDROPHOBIC-AMP2	9	26	

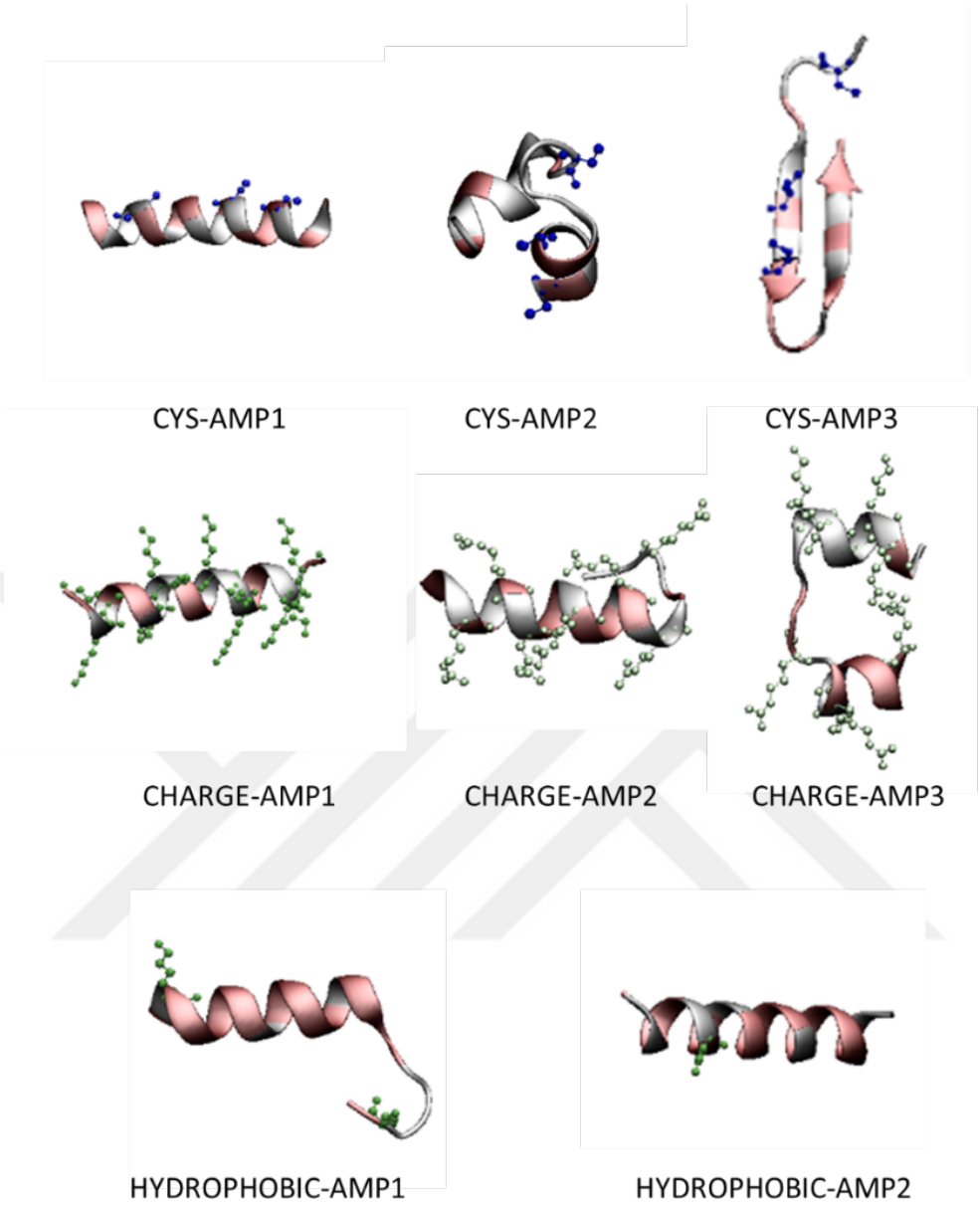


Figure 23. PEPFOLD3 models for the selected peptides (cartoon representation for the best models were used; representations prepared with VMD) Pink: Hydrophobic amino acids, Silver: Hydrophilic amino acids, Blue: Cysteine, Lime: positively charged amino acids.

DISCUSSION

Utilization of state-of-the-art deep learning methodologies to solve antimicrobial peptide recognition problem has been very new. Previous work by Veltri et al. (2018) and many others using the classical machine learning algorithms generally apply these to the well-defined, homogenous datasets comprised mostly from linear cationic peptides such as cathelicidins which are poorly reflecting the diversity of antimicrobial peptides in nature.

Performance results of CNN-BiLSTM model indicate that our model is able to classify and capture antimicrobial peptides based on their properties such as hydrophobicity, chargeness and secondary folds. It is able to classify diverse set of antimicrobial peptides which shows that it is able learn from a rich training set and generalize without overfitting. It proves to be a descent candidate for such tool that can be used in antimicrobial peptide recognition processes to automatize and fasten the candidate selection steps. Further work is needed to perform the final verification of such automatic tool.

Since our model is trained using a heterogenous datasets which consists of large number of diverse subclass of AMPs, our prediction results reflect this notion. That is, the peptides predicted as AMPs with a prediction score of 1 has BlastP hits from a diverse set of microbial organisms such as protozoas, bacteria, fungi and that resulted hits points to several enzymes, transcriptional binding factors etc. which hints to multiple plausible AMP mechanisms for these candidates.

Our dataset is much more challenging. As it can be seen with the performances of various known models, their performance degrades compared to the benchmarking done with their own data set. The key difference is that our dataset contains a greater number of unique antimicrobial peptides which are not limited to a certain sub class of antimicrobial peptides as opposed to current datasets used in similar studies where the main focus mostly on linear cationic peptides. Therefore, our dataset is much more complex and reflects real-world conditions much better than other datasets used in

literature currently. This manifests on the benchmarking performance results where the degrading performance is due to the challenging variety of antimicrobial peptides.

Antimicrobial peptides which are not linear cationic are not well characterized, and that their mechanisms of action along with their unique properties contributing to the antimicrobial activity have not been illuminated completely. We also collected greater number of decoy peptides to be used as the negative class set during the training. We postulate that this also reflects better the real-world conditions of hunting antimicrobial peptides amongst whole known dataset and technically it enables deep learning models to learn better class boundaries without overfitting to such variable set of antimicrobial peptides found in our dataset.

Overall, both CNN based models and BiLSTM models were able to achieve satisfactory performance meanwhile LSTM models were not. We postulate that this is due to short length of peptides which prevents the LSTM models to model the sequence information in one direction. Although, BiLSTM based models and CNN models had similar performance metrics, CNN models had higher MCC scores and precision values. This shows that convolutional layers in CNN models are able to exploit evolutionary and structural information hidden in the sequence by convoluting the data to higher dimensions. These indicatory benchmarks suggested that convolutional layers followed by BiLSTM layers to model the sequence information in high dimensional space. Followed by a dropout layer in order to prevent overfitting, this idea proved useful in that it produced best performance results all around in all metrics except precision which only bested by CNN. This suggest a performance trade-off between recall and MCC with precision where the gain of MCC and recall is much higher than the lost precision score.

The benchmarking results of our models indicate that it has higher AUC, accuracy, recall and precision. For practical purposes, higher recall indicates that our model can find and classify more antimicrobial peptides amongst all possible candidate peptides. Higher precision scores indicate that the peptides classified as antimicrobial by our model has higher probability to be a true antimicrobial peptide. That is, higher precision would reduce unwanted validation costs and wasted time in wet laboratory procedures due to the false positive peptides.

Our results show that our model is able to capture both overall hydrophobicity and net charge of antimicrobial peptides as well as local hydrophobic and highly charged local regions found in their sequence. As it is known that charged regions in the structure allows for selective binding whereas hydrophobic regions in the structure aids in docking to the target protein and thereby enable antimicrobial activity.

It is important to note that despite we removed signal sequences in front of the antimicrobial peptides, our model was able to classify peptide sequences that can be mapped to signal sequences. These AMP candidates are hypothesized to be active given their binding affinity to transcription factors due to their homologous similarities between them and thus have potential to bind to DNA and inhibit paramount functions such as DNA replication, DNA repair along with transcription and translation which would prevent the organism from multiplying, recovering from internal and external DNA stress, and producing vital proteins for various functions.

5 CONCLUSION

Based on the performance results of various models which includes previous SOTA models trained with different datasets, our dataset is much more challenging in that it is harder to achieve close perfect scores in several metrics including MCC and precision. This arise from the fact that our dataset is curated using every non-redundant AMP peptide in the literature which includes both linear cationic and other sub class of AMPs along with several fold of decoy peptides compared to the number of AMPs. As this reflects real world conditions better, it also demonstrates that AMP recognition problem is a challenging problem when taken under realistic conditions. Relatively low score of precision and MCC shows that it is a challenging task to capture true positive amongst masses of decoy peptides which closely resembles the task of hunting AMPs in real world scenarios.

One limitation in this work is the peptides verified to be plausible candidates as predicted by the model have not yet been validated in wet laboratory. Another limitation is that only raw AA sequences have been used to feed and compare the models, more alphabets and sequence input formats can be used to further investigate their effect on performance. In addition, giant neural network models such as transformer-based language models are not included in this research which prevents the SOTA utilization of millions of accumulated peptide sequences to train the classification model.

Future works includes using pre-trained deep learning models such as pre-trained transformers, language models and LSTM based models to increase the performance of the model. On a similar basis, different sequence representation formats such as SMILES and similar protein AA alphabets can be used to increase the classification format. Also, all classification models are trained in a supervised manner in this study which is not ideal in many realistic scenarios. As the annotation efforts have been stumbled behind the data generation models, the DECOY models are actually not “non-AMPs”; they are just not yet proved to be AMPs. For this reason, the most suitable training setting would be PU learning in which the model is trained using verified positive samples alongside with unlabeled samples which can be either a

positive or a negative sample. Last but not least, the most intriguing potential of deep learning applied on this area lies in the automated generation capabilities of such models. It is both reasonable and possibly fruitful choice to invest efforts into training deep generative models that can generate plausible AMP candidates with reasonable performance.

6 REFERENCES

1. Cunha BA. Historical aspects of infectious diseases, part I. Infect Dis Clin. 2004 Mar 1;18(1):xi–xv.

2. Cunha BA. Historical aspects of infectious diseases, part II. Infect Dis Clin. 2004 Jun 1;18(2):xi–xv.

3. Investigation on the Mass Graves Found in House Foundations at the Hamin Mangha Site,Inner Mongolia:Exploration and Reflection on Prehistoric Catastrophic Events--《Archaeology and Cultural Relics》2016年05期 [Internet]. [cited 2021 Apr 6]. Available from: https://en.cnki.com.cn/Article_en/CJFDTotal-KGYW201605005.htm

4. Allen P. THE "JUSTINIANIC" PLAGUE. Byzantion. 1979;49:5–20.

5. Rosen W. Justinian's Flea: The First Great Plague and the End of the Roman Empire. Penguin; 2007. 384 p.

6. Pamuk Ş. The Black Death and the origins of the 'Great Divergence' across Europe, 1300–1600. Eur Rev Econ Hist. 2007 Dec 1;11(3):289–317.

7. Horrox R. The Black Death [Internet]. The Black Death. Manchester University Press; 2013 [cited 2021 Apr 6]. Available from: https://www.manchesterhive.com/view/9781526112712/9781526112712.xml

8. Valleron AJ, Meurisse S, Boelle PY. Historical Analysis of the 1889–1890 Pandemic in Europe. Int J Infect Dis. 2008 Dec 1;12:e95.

9. Valleron A-J, Cori A, Valtat S, Meurisse S, Carrat F, Boëlle P-Y. Transmissibility and geographic spread of the 1889 influenza pandemic. Proc Natl Acad Sci. 2010 May 11;107(19):8778–81.

10. Trilla A, Trilla G, Daer C. The 1918 "Spanish Flu" in Spain. Clin Infect Dis. 2008 Sep 1;47(5):668–73.

11. Fleming A. Penicillin. Br Med J. 1941 Sep 13;2(4210):386.

12. Ventola CL. The Antibiotic Resistance Crisis. Pharm Ther. 2015 Apr;40(4):277–83.

13. Neu HC. The Crisis in Antibiotic Resistance. Science. 1992 Aug 21;257(5073):1064–73.

14. DiMasi JA, Hansen RW, Grabowski HG. The price of innovation: new estimates of drug development costs. J Health Econ. 2003 Mar 1;22(2):151–85.

15. Alanis AJ. Resistance to Antibiotics: Are We in the Post-Antibiotic Era? Arch Med Res. 2005 Nov 1;36(6):697–705.

16. Wang S, Zeng X, Yang Q, Qiao S. Antimicrobial Peptides as Potential Alternatives to Antibiotics in Food Animal Industry. Int J Mol Sci [Internet]. 2016 May 3 [cited 2019 Nov 7];17(5). Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4881439/

17. Savoia D. Plant-derived antimicrobial compounds: alternatives to antibiotics. Future Microbiol. 2012 Aug;7(8):979–90.

18. Huan Y, Kong Q, Mou H, Yi H. Antimicrobial Peptides: Classification, Design, Application and Research Progress in Multiple Fields. Front Microbiol [Internet]. 2020 [cited 2021 Apr 6];11. Available from: https://www.frontiersin.org/articles/10.3389/fmicb.2020.582779/full

19. Malkoski M, Dashper SG, O'Brien-Simpson NM, Talbo GH, Macris M, Cross KJ, et al. Kappacin, a novel antibacterial peptide from bovine milk. Antimicrob Agents Chemother. 2001 Aug;45(8):2309–15.
20. Schittek B, Hipfel R, Sauer B, Bauer J, Kalbacher H, Stevanovic S, et al. Dermcidin: a novel human antibiotic peptide secreted by sweat glands. Nat Immunol. 2001 Dec;2(12):1133–7.
21. Lai Y, Villaruz AE, Li M, Cha DJ, Sturdevant DE, Otto M. The human anionic antimicrobial peptide dermcidin induces proteolytic defence mechanisms in staphylococci. Mol Microbiol. 2007 Jan;63(2):497–506.
22. Seyfi R, Kahaki FA, Ebrahimi T, Montazersaheb S, Eyvazi S, Babaeipour V, et al. Antimicrobial Peptides (AMPs): Roles, Functions and Mechanism of Action. Int J Pept Res Ther. 2020 Sep 1;26(3):1451–63.
23. Pahlavan Y, Kahroba H, Samadi N, Karimi A, Ansarin K, Khabbazi A. Survivin modulatory role in autoimmune and autoinflammatory diseases. J Cell Physiol. 2019 Nov;234(11):19440–50.
24. Zhang L, Shi W, Zeng X-C, Ge F, Yang M, Nie Y, et al. Unique diversity of the venom peptides from the scorpion Androctonus bicolor revealed by transcriptomic and proteomic analysis. J Proteomics. 2015 Oct 14;128:231–50.
25. Wang G, Li X, Wang Z. APD3: the antimicrobial peptide database as a tool for research and education. Nucleic Acids Res. 2016 Jan 4;44(Database issue):D1087–93.
26. Kościuczuk EM, Lisowski P, Jarczak J, Strzałkowska N, Jóźwik A, Horbańczuk J, et al. Cathelicidins: family of antimicrobial peptides. A review. Mol Biol Rep. 2012 Dec 1;39(12):10957–70.
27. Bals R, Wilson JM. Cathelicidins--a family of multifunctional antimicrobial peptides. Cell Mol Life Sci CMLS. 2003 Apr;60(4):711–20.

7 CURRICULUM VITAE