Compilation and Prediction of Hemolytic Peptides using Machine Learning Techniques

Submitted by

Kavin Raj S A
(MT23231)

Under the
Supervision
of

Prof. Gajendra Pal Singh Raghava

Submitted

in partial fulfillment of the requirements for the degree of

Master of Technology

To

Department of Computational Biology,
Indraprastha Institute of Information
Technology, New Delhi

June 2025

Certificate

This is to certify that the thesis titled "Compilation and Prediction of Hemolytic Peptides using Machine Learning Techniques" being submitted by Kavin Raj S A to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

The results contained in this thesis have not been submitted in part or full to any other university or institute for the award of any degree/diploma.

June, 2025

Prof. Gajendra Pal Singh Raghava
Head and Professor
Department of Computational Biology
Indraprastha Institute of Information Technology Delhi
New Delhi 110 020

Acknowledgement

I want to express my heartfelt gratitude and sincere respect to Prof. Gajendra P. S. Raghava from Indraprastha Institute of Information Technology, Delhi, for being my supervisor. His encouragement, mentorship, and insightful guidance fueled the direction of my research and the completion of this thesis.

In addition to my supervisor, I am immensely thankful to Ms. Ayushi Singh and PhD research scholar Mr. Anand Singh Rathore for their constant support, valuable suggestions, and motivation throughout my work.

I would be thankful to the Department of Computational Biology and the Library team at Indraprastha Institute of Information Technology, Delhi, for providing the necessary infrastructure and technical support for my research work.

Lastly, thank my family and friends for their unwavering support, patience, and encouragement throughout this journey. Their belief in me has been a constant source of strength and focus during this endeavor.

Kavin Raj S A

M.Tech Computational Biology

Table of Contents

List of Abbreviations

RCB	Red Blood Cells
SHAP	SHapley Additive exPlanations
LightGBM	Light Gradient Boosting Machine
SMILES	Simplified Molecular Input Line Entry System
MAP	Modification and Annotation in Proteins
PEG	Polyethylene Glycol
PDB	Protein Data Bank
ML	Machine learning
AAC	Amino acid composition
ATC	Atomic Composition
MCC	Matthews Correlation Coefficient
AUC	Area under the curve
XGBoost	Extreme Gradient Boosting
CAMP-R4	Collection of Antimicrobial Peptides
DAMPD	Dragon Antimicrobial Peptide Database
APD	Antimicrobial Peptide Database
GRAVY	Grand Average of hydropathy

List of Figures

List of Tables

Abstract

This thesis describes the compilation, characterization, and prediction of hemolytic peptides, which are responsible for lysing red blood cells. We present Hemolytik2, a comprehensive repository that significantly updates the 2014 Hemolytik database. This new version contains 13,215 entries (7,800 unique peptides), representing a threefold increase over its predecessor, compiled from scientific literature and other peptide databases. Each entry details information such as peptide sequence, terminal modifications, topology, stereochemistry, red blood cell (RBC) source, peptide origin, hemolytic potency, and structural features (SMILES, secondary/tertiary structures). In addition to data compilation, we characterized the peptides and developed a robust method for predicting hemolytic peptides. Peptide features were computed using the widely adopted Pfeature software. A wide range of machine learning techniques, including LightGBM and Random Forest, have been used to develop classification models for discriminating hemolytic and non-hemolytic peptides. SHapley Additive exPlanations (SHAP)-based feature analysis was then applied to identify and rank important features to understand potential of physicochemical descriptors and amino acids. The insights gained from this prediction and feature analysis will be invaluable for the rational design of optimal, safe hemolytic peptides.

Keywords: Hemolytic Peptides, Database, Peptide Toxicity, Feature Analysis, Machine Learning, Therapeutic Peptides, SHAP, SMILES, Peptide Design

Introduction

Peptides are used as therapeutic agents because of their unique properties in regulating biological processes. They have low toxicity and good pharmacokinetic properties and are highly specific to the target [1]. Over the years, peptide screening and synthesis improvements have resulted in the discovery of bioactive functional peptides. These peptides may have uses in cancer treatment, infectious diseases, metabolic conditions, and autoimmune disorders [2,3]. However, despite their potential for treatment, many possible reactive peptides show unwanted toxicity. This toxicity can be divided into immunotoxicity, hemotoxicity, and cytotoxicity [4]. Hemotoxicity involves the breakdown of red blood cells, which leads to serious safety risks and limits its use in clinical research [4]. A hemolytic concentration of 50% (HC50) indicates the level at which 50% of normal human red blood cells break down under normal conditions, and this measure is often used to assess peptide toxicity in experimental setups [5].

Peptides cause hemolytic activity mainly by damaging the membranes of red blood cells [6]. They do this through various mechanisms, including forming pores, displacing lipids, and self-assembling on the surface of membranes [6]. Hemolysis often occurs because of interactions with the phospholipid bilayer of red blood cell membranes, an important factor in the pre-clinic stages of peptide drug development [7]. Predicting or evaluating a peptide's hemolytic potential improves lead compounds and reduces unwanted cytotoxic effects [7]. However, hemolytic toxicity is frequently ignored during early screening. This happens because there are not enough complete and easy-to-access resources available. Most datasets have different purposes and views on research, so necessary information like structural and physicochemical can be missing, which plays an important role in in-depth computational analysis [7].

Peptides with high hemolytic activity mostly tend to be highly active against micro-organisms. However, they can cause hemolysis, the breakdown, and early lysis of red blood cells (RBCs). This lysis releases substances like heme, iron, and hemoglobin into the bloodstream—these released components such as oxidative stress, nitric oxide (NO) scavenging, and increased inflammation. As a result, it leads to serious health issues like atherosclerosis, thrombosis, and kidney injury [8]. Peptides promise as therapeutic agents because of their effect on target selectivity, good environmental tolerability, and lower risk of immune response compared to small organic molecules and antibodies [9].

Historically, finding peptides with low or no hemolytic activity has required much time, effort, and money, especially when evaluating many candidates. Additionally, measuring

hemolytic activity is not standardized [10]. This results in inconsistent outcomes influenced by factors like the source of RBCs (for example, sheep, human, rabbit, or rat), types of buffer, or the amount of DMSO in the experiments, which lead to different structures. A significant challenge in identifying non-hemolytic peptides is the absence of a widely accepted way to measure hemolysis and a clear definition of the minimum hemolytic concentration (MHC) or hemolytic concentration (HC50) [10]. These can vary greatly, with definitions potentially set at 5%, 10%, 50%, or 100% hemolysis. This lack of clarity often leads to the misclassification or poor characterization of peptides, making it challenging to compile reliable "negative" datasets [10].

Hemolytic Database

The fast increase in peptide sequence data from proteomics and peptidomics has created computational methods as a quick and affordable option for predicting hemolytic activity [11]. Several key databases made a significant impact on these hemolytic studies. Hemolytik (2014) is a manually curated database of scientifically experimentally determined hemolytic and non-hemolytic peptides. It compiles data from various sources, including the Antimicrobial Peptide Database(APD) and Swiss-Prot. The database contains about 3,000 entries representing around 2,000 unique peptides [12]. The activity of these peptides is evaluated on RBCs from up to 17 different sources. Database of Antimicrobial Activity and Structure of Peptides (DBAASP) (v2 in 2016, v3 in 2021) is a detailed resource that provides information on antimicrobial, cytotoxic, and hemolytic properties, chemical changes, and 3D structures[13].

UniProt is an informative and broad sequence and annotation database. It offers high-quality, curated information that is used to analyze protein functions. Researchers use it to find proteins, conserved domains, and functional pathways. This database integrates data with multiple genomic and proteomic studies. It includes antimicrobial and toxic peptides, such as hemolytic, crucial for identifying toxic areas in therapeutic peptides [14].DRAMP (Database of Research on Antimicrobial Peptides) is a database for antimicrobial peptides (AMPs), and it offers detailed information on their activity, structure, and source organisms. It also provides information on physicochemical properties and mechanisms of action for AMPs. The database features peptides with activities like hemolytic and cytotoxicity, which helps identify harmful peptides and safe therapeutic candidates. It supports drug development and the design of safer peptide analogs [15].

CAMP_R4 (Collection of Antimicrobial Peptides) is a collection of tested AMPs and contains peptide sequences, structures, and activity classifications. The database provides tools to

predict potential AMPs, and it records curated experimental hemolytic or antimicrobial activity to evaluate peptide toxicity in host cells. This information plays a massive role in peptide design for a better therapeutic index. [16]. APD3 (Antimicrobial Peptide Database version 3) contains natural and synthetic AMPs. It includes antibacterial, antifungal, antiviral, and anticancer peptides and has peptide activity to lyse the antimicrobial. The entries come from various organisms, including humans and microbes like bacteria. The database details peptide length, charge, structure, and hemolysis. APD3 aids in understanding hemolytic activity and is crucial for selecting peptides and creating effective, non-toxic peptide-based treatments [17].

Hemolytic Prediction Tools

Research on computational tools has been suggested for predicting peptide hemolysis. These tools or research methods use machine learning, deep learning techniques, and LLMs (Large language models) to predict hemolytic activity. HemoPI (2016) used Support Vector Machine (SVM) models that used features such as binary profiles, dipeptide composition, motifs, and amino acids [18]. HemoPred (2017) used a Random Forest (RF) classifier based on amino acid and dipeptide composition, as well as physicochemical descriptors [19]. HLPpred-Fuse (2020) uses a two-layer prediction framework method that combines several machine learning classifiers with sequence-based encodings to create probabilistic features [20]. HAPPEN (2020) used artificial neural networks to predict hemolytic activity based only on a simple peptide sequence [21]. HemoPImod (2020) aims to predict the hemolytic activity of chemically modified peptides. It used a random forest model with various correlated features, including 2D and 3D descriptors, fingerprints, and atom/diatom composition [22].

HemoNet (2021) uses a neural network model to analyze the content of amino acids. It also included SMILES-based fingerprint representation for N/C terminal modifications [23]. AMPDeep (2022) applied transfer learning to fine-tune a large transformer-based model (Prot-BERT-BFD) on a small peptide dataset. It successfully leveraged patterns learned from other protein and peptide databases to improve prediction accuracy [24]. EnDL-HemoLyt (2023) combined multiple deep-learning techniques. It used bidirectional long short-term memory, bidirectional temporal convolutional network, and 1D convolutional neural network algorithms, incorporating both handcrafted features and those from deep learning [25]. Hemolytic-Pred (2023) is a machine-learning hemolytic protein predictor using position and composition-based features [26]. Multi-query Similarity Searching (MQSS) Models (2024) introduced a strong model based on complex network science. This model outperformed the best existing machine learning models

in predicting hemolytic activity. MQSS models can identify many more potentially hemolytic peptides than those currently reported, estimating a 3.9-fold increase in actual hemolytic peptides for various endpoints [27]. HemoDL (2024) proposed an ensemble learning model with a double LightGBM framework. This model combines rich sequence-derived and transformer-enhanced information, including features like CTD, BPF, Charge, AAC, GDPC, ATC, and QSO [28]. HemoPI2 (2025) introduced classification and regression models using machine learning and protein language models. These models are designed to identify toxic hemolytic peptides and quantify their hemolytic concentration (HC50). They fill an important gap in current methods, which are often generalized to all vertebrates without specific HC50 prediction abilities [29].

Despite these advancements, ongoing challenges remain. These include the high demand for data from deep learning models, especially when large training datasets are limited. There is also an undercount of hemolytic peptides in current databases. Most methods have previously struggled to predict specific HC50 values for peptides essential for precise drug development. This thesis addresses these key issues by developing and evaluating improved computational models for predicting hemolytic and non-hemolytic peptides. It focuses on accurate classification for effective peptide development in in-silico drug design.

Chapter 2: Compilation of hemolytic peptide

Materials and Methodology

2.1 Data Curation

Hemolytik 2.0 is a comprehensive update to the original Hemolytik database, significantly expanding its coverage of experimentally valid hemolytic and non-hemolytic peptides. A detailed literature survey was conducted on PubMed using the query and removed all the review papers :((((hemolysis) OR (hemolytic)) OR (hemotoxin)) AND (peptide) AND (2013:2024[pdat])) NOT ((((hemolysis) OR (hemolytic)) OR (hemotoxin)) AND (peptide)[14]. Additional data were integrated from well-established peptide repositories such as APD3 (179 peptides), CAMP-R4 (35 peptides), UniProt (268 peptides), and DAMPD (6 peptides) using hemolysis-related keywords. Each selected peptide entry includes detailed annotations: amino acid sequence, origin, terminal modifications, chirality, red blood cell (RBC) source, hemolytic activity (quantified by HC50 or related metrics), and chemical/physical modifications [14].

Database Architecture and Web Interface:

Hemolytik 2.0 was built using a Linux, Apache, MySQL, PHP stack. The backend uses Apache HTTP Server as the web server and MySQL as the relational database management system. Frontend technologies include HTML5, CSS3, JavaScript (v1.8) for responsive design across desktop and mobile platforms, PHP, and PERL for scripting and server-side logic. The overall architecture is shown in Figure 2 and supports secure, fast, and scalable access to hemolytic peptide data. A dynamic, user-friendly interface was designed to provide seamless access to the database contents. The system architecture is illustrated in Figure 2.

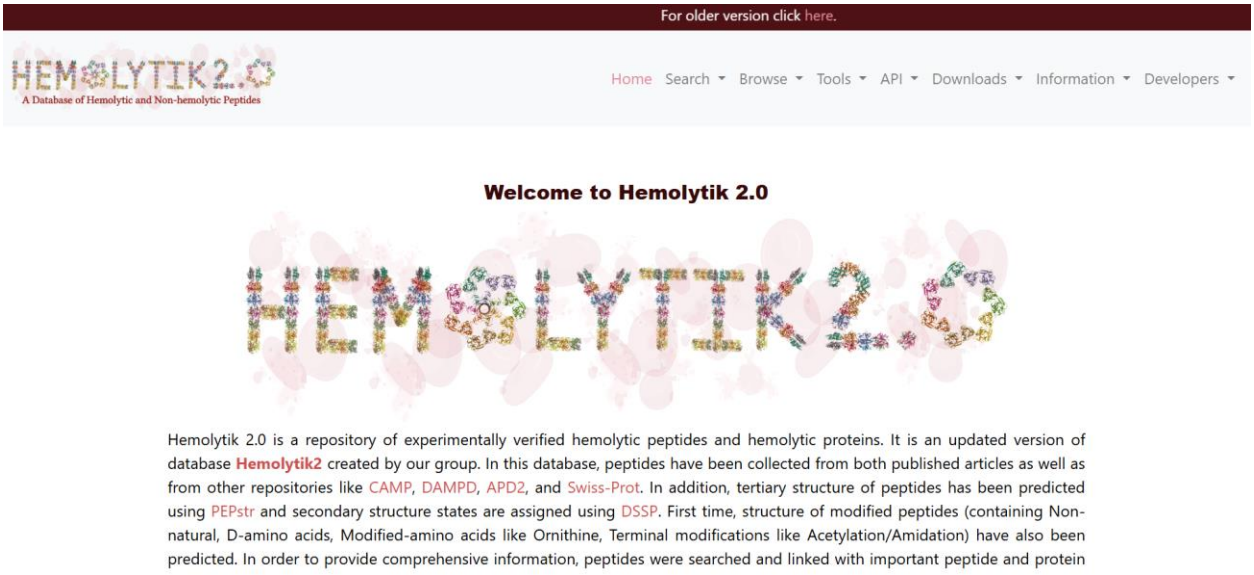

Figure 1: Home page of Hemolytic 2.0

2.3 Data Content of Hemolytik 2.0 Database

Tertiary structures were predicted using PEPstrMOD for peptides >5 residues and I-TASSER for longer sequences. Experimentally determined structures were matched via PDB, and short peptides (<5 residues) were modeled using default backbone dihedrals with energy minimization and MD simulations. DSSP was used to annotate secondary structural features.

2.4 Implementation of Tools:

2.4.1 Search Tool:

Basic Search allows us to do keyword-based searches across fields like peptide name, sequence, chirality, activity, and origin. The supports Boolean logic (AND, OR, NOT) in Advanced Search combines multiple fields. Peptide Search includes both exact match and

subsequence search options. SMILES Search Uses RDKit to support substructure, exact structure, and superstructure queries.

2.4.2 Browsing Tool:

Users can browse entries by Peptide source (human, rat, mouse, etc.), Structure (Linear/Cyclic), Chirality (L, D, or Mix), Biological function (e.g., antimicrobial, anticancer), Terminal and chemical modifications.

2.4.3 Analysis Tool:

BLAST: Supports global alignment for sequence similarity. Smith-Waterman: Optimized for short sequence local alignments. Mapping Tool: Identifies peptide segments in larger sequences. Structure Alignment: Allows 3D structure comparisons using PDB file input.

2.5 Download and API:

Hemolytik 2.0 provides multiple FASTA files of natural and modified peptide sequences. PDB files of predicted 3D structures. Exportable CSV/Excel files for customized queries. Access to all open-access references used for data curation. Hemolytik 2.0 includes a RESTful API to allow us to retrieve data in JSON format programmatically. Querying based on sequence, chirality, structure type, or biological origin. Seamless integration with bioinformatics pipelines and machine learning workflows.

Results and Discussion

Data for Hemolytik 2.0 was gathered and organized through a careful manual curation process from published research studies and several well-known peptide databases. A detailed search was performed in PubMed to find relevant published literature. This search focused on original research articles from 2013 to 2024 that tested the hemolytic activity of peptides using hemolysis assays by different types of RBCs. The specific search query in PubMed extracted 4,533 articles for detailed data extraction. In addition, data came from well-known peptide repositories. This included 35 peptides from CAMP R4, six from DAMPD, 179 from APD3, and 268 from UniProt. We used relevant keywords like "hemolysis" and "hemolytic." We extracted important information for each peptide, including the terminal modifications, amino acid sequence, chirality, biological origin, the specific source of red blood cells (RBCs) used in the assay, and its experimentally determined hemolytic activity.

Organization of Database

The Hemolytik 2.0 database was built using an Apache HTTP Server, with MySQL as its backend. MySQL is an object relational database management system (RDBMS) that helps with data storage and retrieval. The database user interface was created using PHP, CSS, HTML, and JavaScript. PHP and PERL were used for all known gateway and database interface scripts. The choice of MySQL, Apache, and PHP came from their open-source nature and compatibility with different systems. Each entry in the Hemolytik 2.0 database is carefully annotated. This includes the amino acid sequence, peptide name, biological source and origin, functional features, terminal modifications, stereochemistry, and structural classification (linear or cyclic or other structure like bicyclic, branched, stapled). It also contains experimentally verified hemolytic activity data for each peptide. The database provides simple molecular representations of hemolytic peptides in SMILES format, along with predicted tertiary structures and documented secondary structural states. A RESTful API was implemented to improve access to extensive machine-learning applications and drug design tasks. This feature allows the program pipeline to access and automate data retrieval in JSON format. The database supports the MAP format, which enables sequences with modifications at the residue level. These modifications include chemical changes, binding sites, terminal modifications, mutations, and non-standard amino acids.

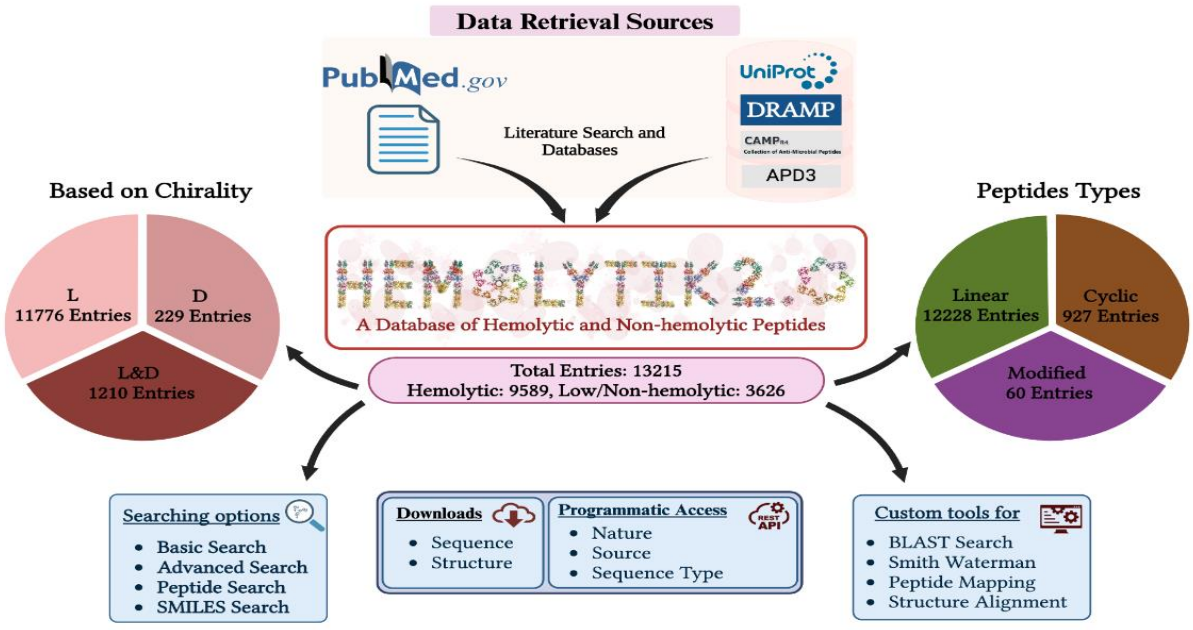

Figure 2: Architecture of Hemolytik 2.0

Database Statistics

Hemolytik 2.0 contains 13,215 curated entries, which include about 8,700 unique peptides. Many of these peptides show additional biological activities, showcasing their multifunctional nature. Specifically, 5,760 peptides have antimicrobial properties, including 2,841 antibacterial, 379 antifungal, and 163 anticancer peptides. The dataset also has 118 peptides with cell penetrating activity and 118 cytotoxic peptides. Around 450 peptide entries currently lack functional annotations or unknown activity.

The hemolytic activity calculated by the hemolytic assay of peptides in the database was tested using red blood cells (RBCs) from 25 different sources. The most common RBC sources are humans (9,255 entries), horses (1,062), sheep (840), and mice (595). The database also includes peptides tested on RBCs from other organisms like pig, porcine, and chicken (298 peptides). Several entries exist for individual peptides, especially when tested at different concentrations, such as mouse and human at 50 μM and 100 μM.

Regarding structural classifications, Hemolytik 2.0 has 12,211 entries for linear peptides and 906 for cyclic peptides. An extra 75 entries represent peptides with complex structures, such as branched, bicyclic, stapled, and macrocyclized forms. For stereochemistry, 11,728 entries contain L-amino acids, 229 include D-amino acids, and 1,210 contain mixed L/D amino acids.

Chemical modifications are well-covered, with 1,935 entries describing changed hemolytic peptides. A broader count shows that 8,340 peptides have one or more chemical modifications meant to improve their bioavailability, resistance to breakdown, or ability to pass through cell membranes. The most common modification is C-terminal amidation, found in 7,189 peptides, followed by N-terminal acetylation in 526 peptides. Both modifications are present in 482 peptides. Other modifications include adding D-amino acids, lipidation, and glycosylation. Structures have been predicted for 10,615 peptides using the PEPstrMOD tool. However, predicting structures was not currently possible for some chemically modified peptides due to limitations in forcefield parameters for non-standard amino acids or modified complex chemical groups, such as 2,3-diaminopropionic acid, norleucine, L-propargylglycine, and ornithine. A subset of 7,654 entries specifically describes natural hemolytic peptides.

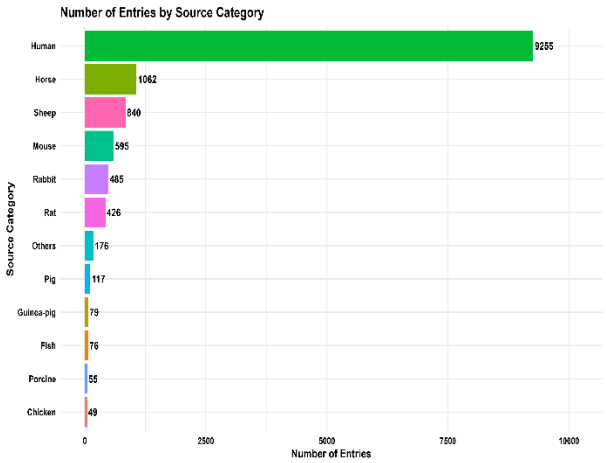

Figure 3: Source of RBCs used to test hemolytic activity

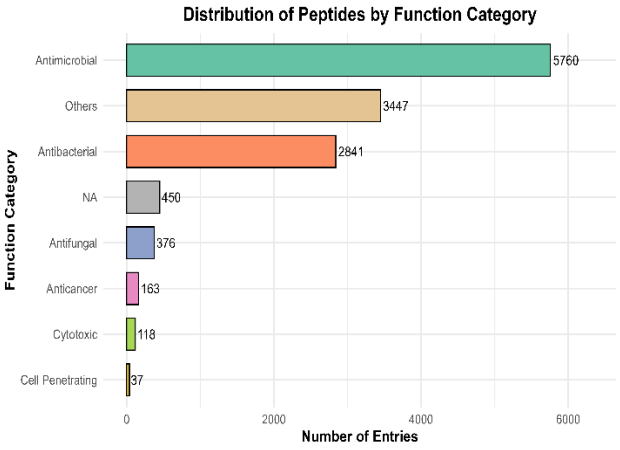

Figure 4: Activity against micro-organisms

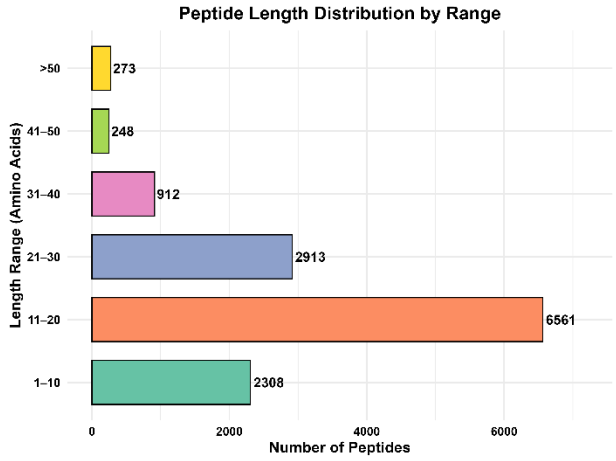

Figure 5: Peptide length

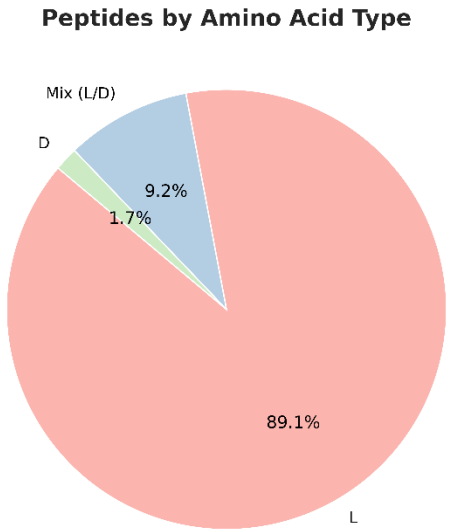

Figure 6: Peptide Amino acid type

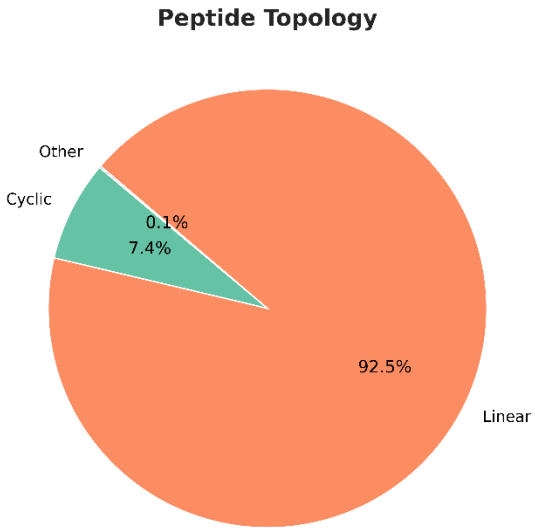

Figure 7: Peptide topology type, here others (Cyclic, Bicyclic, Branched, Stapled, Disulfide Bonding)

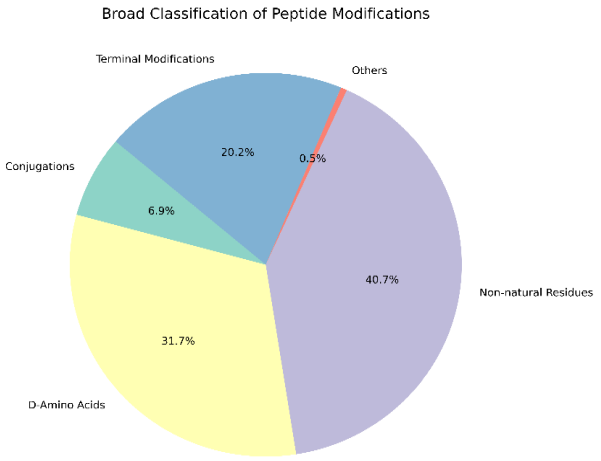

Figure 8: Classification of peptide modification

MAP Format

The MAP format is a modern upgrade to the classic FASTA format, designed to meet the evolving needs of protein research. While FASTA is simple and widely used, it falls short when representing the complexity of real proteins especially those with chemical modifications, non-standard amino acids, or important mutations. MAP solves this by letting scientists add detailed notes directly into the protein sequence using easy-to-read curly brace tags. These tags can highlight modified amino acids, mutation sites, or residues involved in interactions without breaking compatibility with existing software.

MAP also brings structure to the header line, allowing researchers to embed key information about the protein—like its organism, function, or database source—using standardized tags. This means you can see the sequence and what makes the protein unique. Whether you need to track post-translational modifications and engineered mutations or want a richer, more informative protein entry, MAP makes it possible without extra databases or complicated workflows. In short, MAP keeps the simplicity of FASTA but adds the flexibility and detail needed for modern protein science, making it easier for humans and computers to work with complex protein data[31].

Table :1 Types of peptide modification

Modification	**counts**
C-Terminal Modification	117
conjugation	**253**
conjugation, N-Terminal Modification	2
conjugation, non-natural residue	40
D-Amino acids	**840**
D-Amino acids, conjugation	41
D-Amino acids, conjugation, non-natural residue	161
D-Amino acids, conjugation, non-natural residue, N-Terminal Modification	68
D-Amino acids, N-Terminal Modification, C-Terminal Modification	1
D-Amino acids, lipidated	16
D-Amino acids, N-Terminal Modification	23
D-Amino acids, non-natural residue	10
Isobaric residue annotation	6
Isotopic Labelling	9
Mutation	2
N-Terminal Modification	**528**
N-Terminal Modification, C-Terminal Modification	42
non-natural residue	**1450**
non-natural residue, C-Terminal Modification	2
non-natural residue, conjugation, N-Terminal Modification	1
non-natural residue, N-Terminal Modification	30
non-natural residue, N-Terminal Modification, C-Terminal Modification	16
non-natural residue, Non-Natural Modifications (Chemical Derivatization)	1
Protein Variants Deletions + Insertion	1+1
Total modification	**3,661**

Comparison with the Previous Version

Hemolytik 2.0 shows significant improvements over its predecessor. It offers more detailed data and better-quality annotations. The variety of data sources and academic literature has increased notably. The initial database used to collect AMP are UniProt, CAMP, APD2, and DAMPD. Hemolytik 2.0 has added updated and expanded resources like DRAMP 4.0, CAMPR4, UniProt and APD3. The number of articles reviewed by PubMed during the initial filtering jumped from about 900 to 4,350. Also, the number of experimentally data manually curated publications related to peptide information increased from 446 to 1,228.

Regarding data volume, the first version of Hemolytik had 2,970 peptide entries collected successfully, while Hemolytik 2.0 has 13,215 curated entries. The total of unique hemolytic peptides rose from 1,750 to 5,598. Similarly, entries for non-hemolytic peptides grew from 295 to 2,410. The range of red blood cell (RBC) sources used in hemolysis assays increased from 17 species in the first version to 20 in this update.

Hemolytik 2.0 offers better classification for different peptide types. The original version mainly focused on linear peptides comprising 2,669 entries and 301 cyclic peptides. The updated version now includes 12,219 linear peptides, 912 cyclic peptides, and 84 peptides with more complex structures, such as branched, stapled, bicyclic, and macrocyclic forms. It also includes a greater variety of peptide stereochemistry. The number of entries for L-amino acids grew from 2,583 to 11,751. At the same time, D-amino acids rose to 225, and the section for peptides with mixed stereochemistry increased to 1,215.

The latest version provides better information on chemical modifications. The number of entries for chemically modified peptides has increased dramatically from 221 to 2,548. Much research has focused on this area, especially terminal and non-residual modifications. The previous version had little information, listing only 9 types of C-terminals and 25 types of N-terminal modifications. In comparison, Hemolytik 2.0 now describes 199 kinds of N-terminal modifications and 51 kinds of C-terminal modifications, including lipidation, acetylation, amidation, and various other structural changes.

Table 2: Comparison with previous version

S.No.	Study	Hemolytik (paper/web)	Hemolytik 2.0
1	No. of entries	2970	13,215
2	No. of Unique Hemolytic peptide	1750	5326
3	No. of Unique Non-Hemolytic peptide	295	1376
4	No. of Unique low-Hemolytic peptide	0	862
5	Databases used	APD2, CAMP, Uniprot, DAMPD	APD3, CAMPR4, Uniprot, DRAMP 4.0
6	Pubmed paper referred	~900	4350
7	No. of paper used to extracted information	446	1228
8	RCB source	17	22
9	non- hemolytic entries	319	2,258
10	Linear	2669	12,219
11	Cyclic	301	912
12	Branched, Bicyclic, Stapled, Macrocyclized	-	84
13	only L	2583	9,018
14	only D	47	176
15	Mix	340	837
16	Chemical modification	237	400
17	Cter modification including Amidation, acetylation	9	45
18	Nter modification including Amidation, acetylation	17	163

Comparison with Other Databases

Hemolytik 2.0 brings together information from several well-known peptide databases, including the APD, UniProt, CAMP-R4, and the DAMPD. The manuscript points out that most existing datasets are scattered across different publications or do not have the structural and physicochemical details needed for effective computational analyses. The first version of Hemolytik was recognized as the first extensive database to provide experimental data on hemolytic peptides and their strengths. The updated Hemolytik 2.0 builds on this by adding data from newer and more extensive sources, such as DRAMP 4.0, APD3, CAMPR4 and UniProt and to create a more unified and thoroughly detailed platform for researchers.

Chapter 3: Prediction of hemolytic peptide

Materials and Methods

Feature generation

We generated nearly 9000 features by using a bioinformatic tool called Pfeature. We calculated 8976 features, which were used for positional analysis, structural analysis, and model predictions. The details and length of each feature are listed in Table 3.

Table 3: List of all the generated features and its length

S.No.	PFeature Name	Dimension/Count
1	Shannon-Entropy of Protein (SEP)	1
2	Sequence order coupling number (SOC)	2
3	Bond composition (BTC)	4
4	Atom composition (ATC)	5
5	Amino acid composition (AAC)	20
6	Distance distribution of residue (DDR)	20
7	Residue Repeat Information (RRI)	20
8	Shannon Entropy of Residues (SER)	20
9	Pseudo amino acid composition (PAAC)	21
10	Amphiphilic pseudo amino acid composition (APAAC)	23
11	Shannon entropy of physico-chemical properties (SPC)	25
12	Physicochemical Properties Composition (PCP)	30

13	Quasi-sequence order (QSO)	42
14	Conjoint Triad Calculation (CTC)	343
15	Dipeptide composition (DPC)	400
16	Tripeptide composition (TPC)	8000

The Pfeature computational tool allows for the analysis of peptide sequence and physicochemical properties. It provides insights into the composition, arrangement, and distribution of amino acids and the different structural and chemical properties of peptides and proteins. ProteinAnalysis identifies traits like peptide length, isoelectric point (pI), molecular weight, instability index, GRAVY, and extinction coefficient (ExtCoeff). It integrates these properties, including sequence patterns and physicochemical traits relevant for predicting hemolytic and non-hemolytic effects through a machine learning classification model.

Five-fold cross validation

To make sure our machine learning models were not biased by repeated related values or overfitting, used a 5-fold cross-validation method. By splitting the dataset into training and validation in an 80:20 ratio. We evaluated the machine learning classification models using 5-fold cross validation on 80% of the training data while keeping the remaining 20% separated from the original. This machine learning method divides the training data into five equal parts to reduce overfitting. We used four parts for training and set the last part as a test set for internal evaluation. We repeated this process 5 times, ensuring that each part was tested at once, and finally calculated the mean[31].

Evaluation parameters

To evaluate how well machine learning classification models work, we used standard ML assessment metrics like specificity (spec), accuracy (acc), sensitivity (sen), MCC, and AUC. Sensitivity (Equation 1) measures the model's ability to identify hemolytic activity, while specificity (Equation 2) determines how well the model predicts hemolytic activity. Accuracy shows the proportion of correctness of the predictions for hemolytic and non-hemolytic cases (Equation 3). The MCC indicates the relationship between predicted outcomes and actual values given in the original dataset (Equation 4). The AUC provides a measure that does not depend on a threshold; it shows how effectively the model distinguishes between the two classes, with sensitivity on one axis and specificity on the other. [32].

Sensitivity (Recall):

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{i}$$

Specificity:

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{ii}$$

Accuracy:

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

Matthews Correlation Coefficient (MCC):

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

Where True negative (TN), False negative (FN), True positive (TP), and False positive (FP).

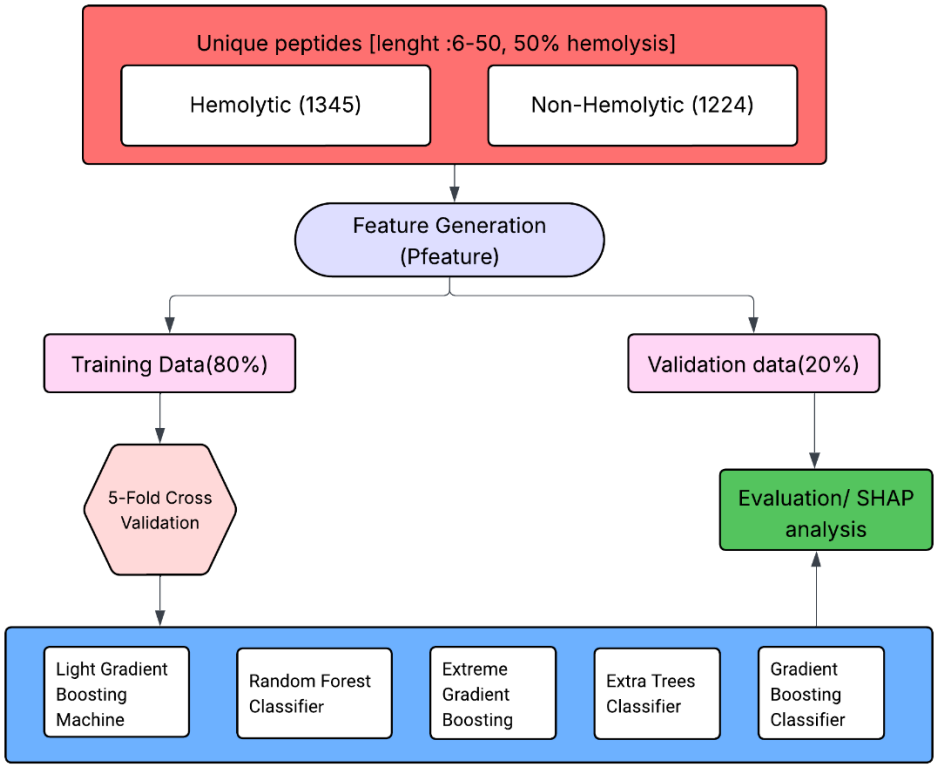

Figure 9: Structural overview of the predictive model for hemolytic peptide

Dataset compilation and preprocessing

The Hemolytik 2.0 database provided 13,215 hemolytic peptide sequences. We focused only on natural peptides and removed any duplicate sequences. Based on prior studies, we found that the best length for hemopi2 data was between 6 to 50 amino acids, with a hemolysis rate of 50%. To ensure comparability, we looked at the amino acid composition of the non-hemolytic peptide negative dataset to confirm that it matched the average composition of naturally occurring proteins. In the end, we created a positive dataset hemolytic with 1,345 hemolytic peptides and a negative dataset with 1,224 non-hemolytic peptides, ensuring all are unique peptide entries between the two sets.

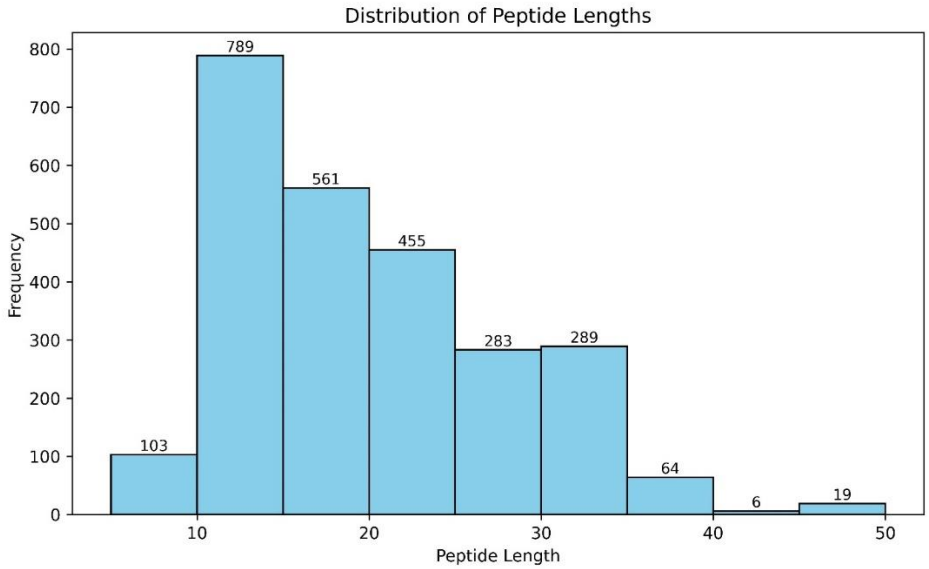

Figure 10: Peptide length distribution of selected peptides

Results and Discussion

Compositional analysis

The AAC of hemolytic peptide was calculated, and it was found that hemolytic sequences tend to have higher levels of lysine, leucine, glycine, arginine, and proline found in nature peptide sequence, as shown in Figure 11.

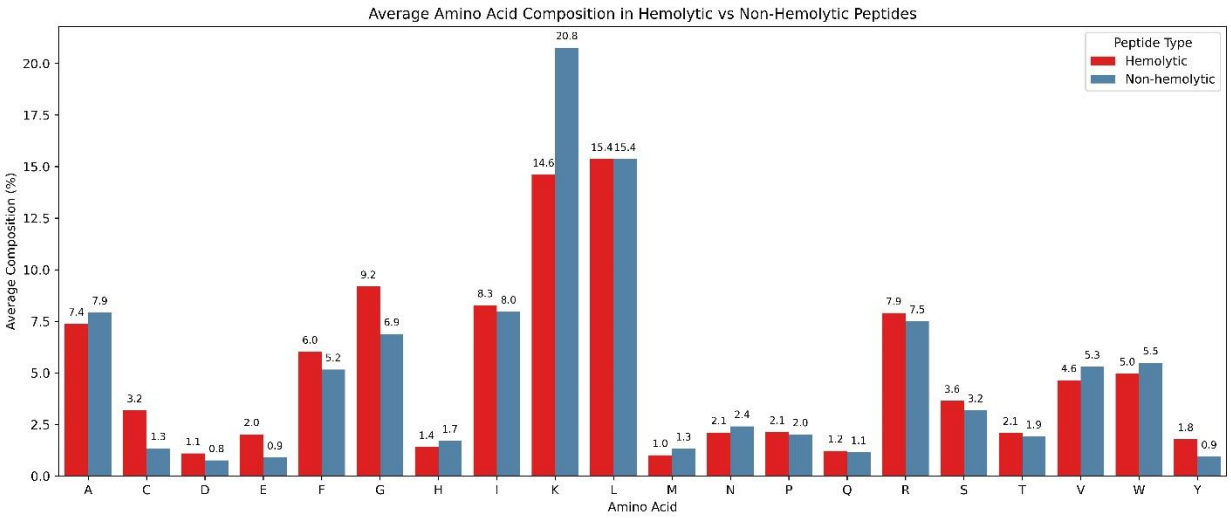

Figure 11: ACC of Hemolytic and Non-Hemolytic sequences

Table 4: Welch's T-Test analysis with respect to hemolytic and non-hemolytic

#Statistical Significance (p < 0.05)

Amino Acids	mean_hemo_%	mean_nonhemo_%	difference_%	p_value
K	14.601	20.759	-6.158	0
G	9.183	6.881	2.302	0
C	3.179	1.341	1.838	0
E	2.021	0.9	1.121	0
F	6.022	5.163	0.859	0.002
Y	1.784	0.943	0.841	0
V	4.64	5.298	-0.657	0.009
A	7.366	7.92	-0.554	0.115
W	4.969	5.492	-0.523	0.17
S	3.642	3.189	0.454	0.018
R	7.901	7.505	0.396	0.416
D	1.094	0.757	0.337	0

M	1.01	1.333	-0.323	0.005
I	8.273	7.964	0.309	0.305
N	2.105	2.408	-0.302	0.032
H	1.403	1.702	-0.299	0.056
T	2.088	1.923	0.165	0.258
P	2.133	2.015	0.118	0.483
Q	1.22	1.148	0.073	0.499
L	15.36	15.358	0.002	0.997

Based on Welch's t-test analysis comparing the amino acid composition of hemolytic and non-hemolytic peptides, the following significant findings were made:

Hemolytic peptides have a higher content of Glycine (G) (+2.30%, p = 0.000), Cysteine (C) (+1.84%, p = 0.000), Glutamic acid (E) (+1.12%, p = 0.000), Tyrosine (Y) (+0.84%, p = 0.000), and Phenylalanine (F) (+0.86%, p = 0.002). This suggests that these amino acids may play a role in hemolytic activity.

On the other hand, non-hemolytic peptides are richer in Lysine (K) (−6.16%, p = 0.000), Valine (V) (−0.66%, p = 0.009), and Methionine (M) (−0.32%, p = 0.005). While Alanine (A) (−0.55%, p = 0.115) and Tryptophan (W) (−0.52%, p = 0.170) show reduced levels in hemolytic peptides, these differences are not significant.

This difference in composition highlights the unique amino acid preferences between hemolytic and non-hemolytic peptides and points out possible sequence features that relate to hemolytic function.

SHAP Analysis

SHAP analysis is a feature explainatiory method used to explain the prediction output of machine learning models, attributing each feature's contribution to a specific prediction. It helps understand why a model made a specific prediction by assigning Shapley values to each feature. These values indicate how much each feature influenced the prediction compared to the average prediction.

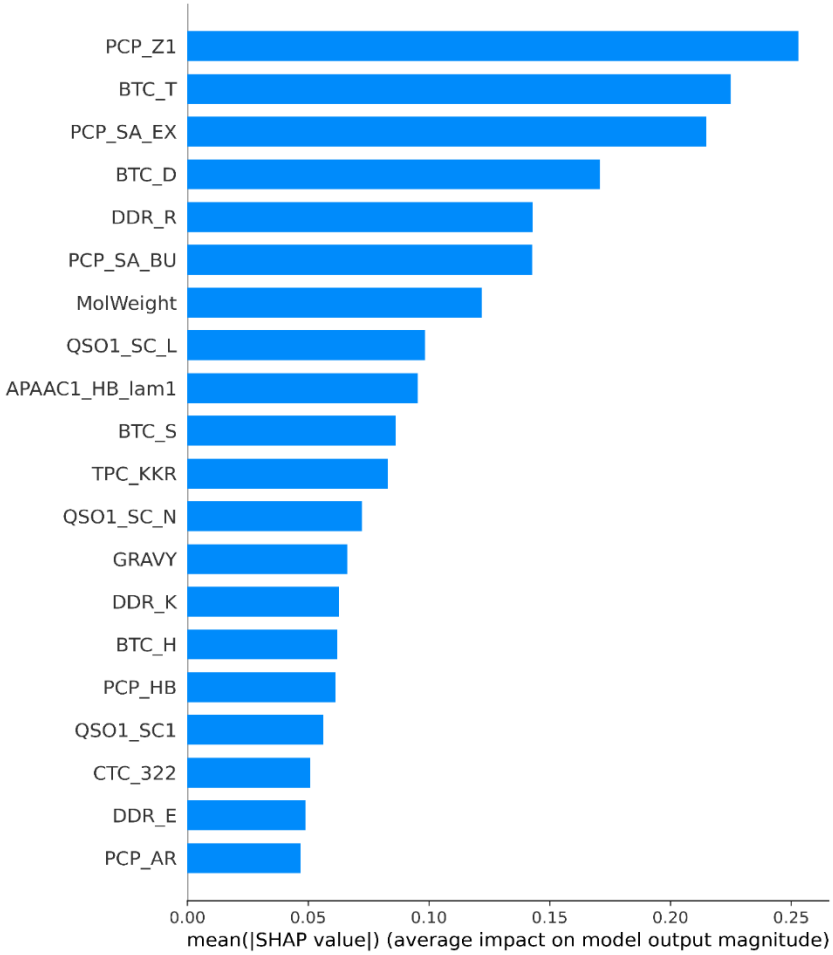

Figure 12: Top 20 Feature impact on complete (Pfeature + ProteinAnalysis) prediction

Based on the SHAP Analysis Figure 12, summary bar plot, and the corresponding feature descriptions, here is a clear interpretation of the top and bottom features influencing your model for predicting hemolytic peptides:

Top Contributing Features

These features have the highest mean absolute SHAP values, indicating they have the most influence on the model's prediction:

PCP_Z1 – High importance suggests that advanced physicochemical properties like electronic characteristics, steric hindrance, or molecular volume significantly contribute to distinguishing hemolytic peptides.

BTC_T (Bond Type Composition - Total Bonds) – Indicates that bond composition patterns are strongly associated with peptide behavior, possibly reflecting structural rigidity or flexibility.

PCP_SA_EX (Solvent Accessibility - Exposed residues) – Suggests hemolytic activity is influenced by surface exposure of residues, potentially relevant for membrane interaction.

BTC_D (Double Bonds) – Structural features involving double bonds may affect the conformation or reactivity of the peptide.

DDR_R (Distance Distribution of Repeats - Arginine) – Implies that the spatial pattern of Arginine residues (positively charged) is crucial, which aligns with Arginine's known role in membrane disruption.

Least Contributing Features: These had the lowest SHAP values, meaning they had minimal individual impact on the model output:

PCP_HB (Hydrophobicity) – Surprisingly, it caused a low impact on the prediction, suggesting overall hydrophobicity alone may not differentiate hemolytic activity well in your dataset.

QSO1_SC1 – Quasi-sequence order (side chain interactions) contributes very little, perhaps due to redundancy with other features.

CTC_322 – One specific Composition/Transition/Distribution index, likely redundant or irrelevant for hemolytic behavior.

DDR_E (Glutamic Acid Repeats) – Indicates spatial repetition of Glutamic Acid (negatively charged) is not a strong determinant.

PCP_AR (Aromatic residue composition) – Suggests aromaticity alone does not strongly influence hemolysis, possibly because it interacts with other factors like charge or sequence context.

The model heavily relies on physicochemical properties, structural features, and charged residue patterns, particularly surface exposure and Arginine distribution. Traditional assumptions like hydrophobicity or aromaticity play a more limited role individually, though they may still matter in combination.

Important Pfeature:

The important Pfeature are isolated by using the name of each top 20 feature and made a mean sum of the feature in the same group.

Important rank	Group	SHAP value
1	PCP	1.098393

2	DPC	0.628656
3	DDR	0.552236
4	BTC	0.544091
5	QSO	0.464621
6	CTC	0.441224
7	TPC	0.290788
8	SEP	0.22915
9	AAC	0.195343
10	APAAC	0.165467

Table 5: Top 10 Pfeatures in feature grouping

Machine Learning Classification

Result of pfeature + length+ Molecular Weight + Instability Index+ gravy (Grand Average of Hydropathy) + ExtCoeff (Extinction Coefficient) + pI (Isoelectric Point)

		Training dataset					**Testing dataset**				
Model	**Time (s)**	**Sens(%)**	**Spec(%)**	**Acc(%)**	**AUC**	**MCC**	**Sens(%)**	**Spec(%)**	**Acc(%)**	**AUC**	**MCC**
LightGBM	**19.61**	**80.568**	**76.504**	**78.638**	**0.8663**	**0.57208**	**80.67**	**80.41**	**80.54**	**0.8699**	**0.6104**
RandomForest	16.52	77.968	78.036	78.008	0.85942	0.5602	77.7	79.59	78.6	0.8579	0.5723
ExtraTrees	33.89	78.714	75.992	77.422	0.85662	0.54782	77.32	80.41	78.79	0.8485	0.5767
XGBoost	61.61	79.27	76.096	77.762	0.86114	0.55426	79.93	81.33	80.74	0.8755	0.601
GradientBoosting	359.54	77.78	75.992	76.936	0.84508	0.5386	78.44	78.78	78.6	0.847	0.5717

Table 6: Result of Pfeature + ProteinAnalysis

Light Gradient Boosting Machine (LightGBM) showed consistently strong performance:

Test Accuracy: 80.54%, Test AUC: 0.8699, Test MCC: 0.6104; this confirms that Pfeature and Protein Analysis play a huge role in predicting hemolytic and non-hemolytic peptides.

Top 10 SHAP Pfeature

		Training dataset					Testing dataset				
Model	**Time (s)**	**Sens(%)**	**Spec(%)**	**Acc(%)**	**AUC**	**MCC**	**Sens(%)**	**Spec(%)**	**Acc(%)**	**AUC**	**MCC**
LightGBM	**15.33**	**80.202**	**76.604**	**78.492**	**0.8606**	**0.56892**	**79.55**	**80.82**	**80.16**	**0.8653**	**0.6031**
RandomForest	17.12	78.246	77.628	77.956	0.8608	0.55892	75.84	77.96	76.85	0.8472	0.5374
ExtraTrees	33.85	78.246	76.3	77.324	0.85076	0.54588	76.95	77.55	77.24	0.8429	0.5445
XGBoost	45.9	79.738	76.502	78.2	0.86174	0.5635	77.7	80.41	78.99	0.8643	0.5804
GradientBoosting	302.42	78.712	76.198	77.52	0.8424	0.54968	77.32	76.73	77.04	0.8431	0.5403

Table 7: Prediction result of Top 10 PFeature

Light Gradient Boosting Machine (LightGBM) showed consistently strong performance:

Highest Accuracy (80.16%), Highest AUC (0.8653), Highest MCC (0.6031), Balanced (Sensitivity 79.55%, Specificity 80.82%).

Top 5 SHAP Pfeature prediction

		Training dataset					Testing dataset				
Model	**Time (s)**	**Sens(%)**	**Spec(%)**	**Acc(%)**	**AUC**	**MCC**	**Sens(%)**	**Spec(%)**	**Acc(%)**	**AUC**	**MCC**
LightGBM	4.76	79.642	75.992	77.91	0.8641	0.5573	79.55	80.41	79.96	0.8738	0.5991
RandomForest	6.42	81.126	78.852	80.05	0.86638	0.6009	79.55	80	79.77	0.8635	0.5951
ExtraTrees	5.12	79.456	76.5	78.056	0.85974	0.56076	77.32	76.73	77.04	0.8484	0.5403
XGBoost	**9.99**	**80.292**	**75.378**	**77.956**	**0.86036**	**0.5585**	**81.41**	**79.18**	**80.35**	**0.8698**	**0.6061**

GradientBoosting	28.99	78.244	74.46	76.45	0.84502	0.52822	76.95	75.92	76.46	0.8441	0.5284

Table 8: Prediction result of Top 5 PFeature

Extreme Gradient Boosting (XGBoost): Highest Sensitivity (81.41%), Highest Accuracy (80.35%), MCC (0.6061), AUC (0.8698) is slightly less than LightGBM.

Result of PCP only prediction

		Training dataset					**Testing dataset**				
Model	**Time (s)**	**Sens(%)**	**Spec(%)**	**Acc(%)**	**AUC**	**MCC**	**Sens(%)**	**Spec(%)**	**Acc(%)**	**AUC**	**MCC**
LightGBM	1.41	74.534	73.028	73.82	0.81508	0.4759	73.61	75.51	74.51	0.8147	0.4906
RandomForest	8.33	74.064	74.66	74.354	0.81972	0.488	71	76.73	73.74	0.8197	0.4772
ExtraTrees	2.23	74.344	73.438	73.92	0.81326	0.47798	73.23	74.29	73.74	0.811	0.4747
XGBoost	2.15	72.484	71.498	72.022	0.80298	0.44046	74.72	74.29	74.51	0.8118	0.4897
GradientBoosting	7.71	72.766	70.78	71.826	0.79078	0.43578	70.63	71.43	71.01	0.7745	0.4202

Table 9: Prediction result of PCP PFeature

The PCP-only approach resulted in significantly lower predictive performance, indicating that while PCP features carry some biological relevance, they are insufficient on their own for accurate classification.

Chapter 4: Conclusion

In this thesis, we developed a solid database focused on hemolytic peptides. The study aims to explore and predict their hemolytic potential. The dataset includes hemolytic and non-hemolytic peptides with verified activity profiles and detailed physicochemical features. The analysis of the amino acid composition and the use of machine learning (ML) models to find patterns between hemolytic and non-hemolytic sequences. Among the models we tested, LightGBM (Light Gradient Boosting Machine) consistently performed best in key evaluation metrics, including accuracy,

Area Under the ROC Curve (AUC), and Matthews Correlation Coefficient (MCC). Its ability to manage high-dimensional feature spaces and capture complex nonlinear relationships made it the most trustworthy option for this binary classification task by using SHAP (Shapley Additive Explanations) analysis to make the model easier to understand. SHAP values helped us pinpoint the most important features affecting hemolytic predictions. The top-ranked features included PCP_Z1 (the first principal component of physicochemical properties), BTC_T (topological descriptor), and PCP_SA_EX (solvent accessibility property), all of which strongly indicated hemolytic activity.

In contrast, features such as GRAVY (grand average of hydropathy), PCP_AR (aromaticity-related descriptor), and DDR_E (electrostatic descriptor) were mainly linked to non-hemolytic peptides. The reduced features are based on SHAP rankings to assess the redundancy and importance of features. Models trained using only the Top 10 and Top 5 SHAP-ranked features showed a minimal drop in performance, suggesting that these key features capture the most informative aspects of hemolytic behavior. However, a model trained solely on PCP (physicochemical properties) did not reach this level of performance, indicating that PCP features alone cannot fully capture the complexity of hemolytic behavior. In the future, we can implement prediction analysis using deep learning and LLM. Also, feature preparation and prediction of the modified sequence will be innovative.

Bibliography

1. Rathore, Anand Singh, Nishant Kumar, Shubham Choudhury, Naman Kumar Mehta, and Gajendra PS Raghava. "Prediction of hemolytic peptides and their hemolytic concentration." *Communications Biology* 8, no. 1 (2025): 176.
2. Gautam, Ankur, Kumardeep Chaudhary, Sandeep Singh, Anshika Joshi, Priya Anand, Abhishek Tuknait, Deepika Mathur, Grish C. Varshney, and Gajendra PS Raghava. "Hemolytik: a database of experimentally determined hemolytic and non-hemolytic peptides." *Nucleic acids research* 42, no. D1 (2014): D444-D449.
3. Mehta, Naman Kumar, Anjali Lathwal, Rajesh Kumar, Dilraj Kaur, and Gajendra PS Raghava. "In silico tool for Predicting, Designing and Scanning IL-2 inducing peptides." bioRxiv (2021): 2021-06.
4. Kumar, Vinod, Rajesh Kumar, Piyush Agrawal, Sumeet Patiyal, and Gajendra PS Raghava. "A method for predicting hemolytic potency of chemically modified peptides from its structure." *Frontiers in pharmacology* 11 (2020): 54.
5. Plisson, Fabien, Obed Ramírez-Sánchez, and Cristina Martínez-Hernández. "Machine learning-guided discovery and design of non-hemolytic peptides." *Scientific reports* 10, no. 1 (2020): 16581.
6. Win, Thet Su, Aijaz Ahmad Malik, Virapong Prachayasittikul, Jarl E. S Wikberg, Chanin Nantasenamat, and Watshara Shoombuatong. "HemoPred: a web server for predicting the hemolytic activity of peptides." *Future medicinal chemistry* 9, no. 3 (2017): 275-291.
7. Salem, Milad, Arash Keshavarzi Arshadi, and Jiann Shiun Yuan. "AMPDeep: hemolytic activity prediction of antimicrobial peptides using transfer learning." *BMC bioinformatics* 23, no. 1 (2022): 389.
8. Sharma, Ritesh, Sameer Shrivastava, Sanjay Kumar Singh, Abhinav Kumar, Amit Kumar Singh, and Sonal Saxena. "EnDL-HemoLyt: ensemble deep learning-based tool for identifying therapeutic peptides with low hemolytic activity." *IEEE Journal of Biomedical and Health Informatics* (2023).
9. Karasev, Dmitry A., Georgii S. Malakhov, and Boris N. Sobolev. "Quantitative prediction of hemolytic activity of peptides." Computational Toxicology 32 (2024): 100335.
10. Chaudhary, Kumardeep, Ritesh Kumar, Sandeep Singh, Abhishek Tuknait, Ankur Gautam, Deepika Mathur, Priya Anand, Grish C. Varshney, and Gajendra PS Raghava. "A web server and mobile app for computing hemolytic potency of peptides." *Scientific reports* 6, no. 1 (2016): 22843.
11. Yaseen, Adiba, Sadaf Gull, Naeem Akhtar, Imran Amin, and Fayyaz Minhas. "HemoNet: Predicting hemolytic activity of peptides with integrated feature learning." Journal of bioinformatics and computational biology 19, no. 05 (2021): 2150021.
12. Timmons, Patrick Brendan, and Chandralal M. Hewage. "HAPPENN is a novel tool for hemolytic activity prediction for therapeutic peptides which employs neural networks." *Scientific reports* 10, no. 1 (2020): 10869.

13. Raza, Ali, and Hafiz Saud Arshad. "Prediction of Hemolysis Tendency of Peptides using a Reliable Evaluation Method." *arXiv preprint arXiv:2012.06470* (2020).
14. Hasan, Md Mehedi, Nalini Schaduangrat, Shaherin Basith, Gwang Lee, Watshara Shoombuatong, and Balachandran Manavalan. "HLPpred-Fuse: improved and robust prediction of hemolytic peptide and its activity by fusing multiple feature representation." *Bioinformatics* 36, no. 11 (2020): 3350-3356.
15. Singh, Ayushi, Kavin Raj SA SA, Anand Singh Rathore, and Gajendra PS Raghava. "Hemolytik2: An Updated Database of Hemolytic Peptides and Proteins." *bioRxiv* (2025): 2025-05.
16. Zhao, Ya, Shengli Zhang, and Yunyun Liang. "HemoFuse: multi-feature fusion based on multi-head cross-attention for identification of hemolytic peptides." *Scientific Reports* 14, no. 1 (2024): 22518.
17. Abdelbaky, Ibrahim, Mohamed Elhakeem, Hilal Tayara, Elsayed Badr, and Mustafa Abdul Salam. "Enhanced prediction of hemolytic activity in antimicrobial peptides using deep learning-based sequence analysis." *BMC bioinformatics* 25, no. 1 (2024): 368.
18. Capecchi, Alice, Xingguang Cai, Hippolyte Personne, Thilo Köhler, Christian van Delden, and Jean-Louis Reymond. "Machine learning designs non-hemolytic antimicrobial peptides." *Chemical science* 12, no. 26 (2021): 9221-9232.
19. Capecchi, Alice, Xingguang Cai, Hippolyte Personne, Thilo Köhler, Christian van Delden, and Jean-Louis Reymond. "Machine learning designs non-hemolytic antimicrobial peptides." *Chemical science* 12, no. 26 (2021): 9221-9232.
20. Wang, Guangshun. "The antimicrobial peptide database is 20 years old: recent developments and future directions." *Protein Science* 32, no. 10 (2023): e4778.
21. Perveen, Gulnaz, Fahad Alturise, Tamim Alkhalifah, and Yaser Daanial Khan. "Hemolytic-Pred: a machine learning-based predictor for hemolytic proteins using position and composition-based features." *Digital Health* 9 (2023): 20552076231180739.
22. Yang, Sen, and Piao Xu. "HemoDL: Hemolytic peptides prediction by double ensemble engines from Rich sequence-derived and transformer-enhanced information." *Analytical Biochemistry* 690 (2024): 115523.
23. Castillo-Mendieta, Kevin, Guillermin Agüero-Chapin, Edgar Marquez, Yunierkis Perez-Castillo, Stephen J. Barigye, Mariela Pérez-Cárdenas, Facundo Peréz-Giménez, and Yovani Marrero-Ponce. "Multiquery similarity searching models: an alternative approach for predicting hemolytic activity from peptide sequence." *Chemical Research in Toxicology* 37, no. 4 (2024): 580-589.
24. Zakharova, Elena, Markus Orsi, Alice Capecchi, and Jean-Louis Reymond. "Machine learning guided discovery of non-hemolytic membrane disruptive anticancer peptides." *ChemMedChem* 17, no. 17 (2022): e202200291.
25. Bhatnagar, Pranshul, Yashi Khandelwal, Shagun Mishra, Arnab Dutta, Debirupa Mitra, and Swati Biswas. "Predicting antibacterial activity, efficacy, and hemotoxicity of peptides using an explainable machine learning framework." *Process Biochemistry* 145 (2024): 163-174.

26. Santos-Junior, Célio Dias, Shaojun Pan, Xing-Ming Zhao, and Luis Pedro Coelho. "Macrel: antimicrobial peptide screening in genomes and metagenomes." *PeerJ* 8 (2020): e10555.
27. Almotairi, Sultan, Elsayed Badr, Ibrahim Abdelbaky, Mohamed Elhakeem, and Mustafa Abdul Salam. "Hybrid transformer-CNN model for accurate prediction of peptide hemolytic potential." *Scientific Reports* 14, no. 1 (2024): 14263.
28. Bin Hafeez, Ahmer, Xukai Jiang, Phillip J. Bergen, and Yan Zhu. "Antimicrobial peptides: an update on classifications and databases." *International journal of molecular sciences* 22, no. 21 (2021): 11691.
29. Rathore, Anand Singh, Nishant Kumar, Shubham Choudhury, Naman Kumar Mehta, and Gajendra PS Raghava. "Prediction of hemolytic peptides and their hemolytic concentration (HC50)." bioRxiv (2024): 2024-07.
30. Tsai, Cheng-Ting, Chia-Wei Lin, Gen-Lin Ye, Shao-Chi Wu, Philip Yao, Ching-Ting Lin, Lei Wan, and Hui-Hsu Gavin Tsai. "Accelerating antimicrobial peptide discovery for who priority pathogens through predictive and interpretable machine learning models." *ACS omega* 9, no. 8 (2024): 9357-9374.
31. Shendre, Akshay, Naman Kumar Mehta, Anand Singh Rathore, Nishant Kumar, Sumeet Patiyal, and Gajendra PS Raghava. "MAP Format for Representing Chemical Modifications, Annotations, and Mutations in Protein Sequences: An Extension of the FASTA Format." *arXiv preprint arXiv:2505.03403* (2025).
32. Khabbaz, Hossein, Mohammad Hossein Karimi-Jafari, Ali Akbar Saboury, and Bagher BabaAli. "Prediction of antimicrobial peptides toxicity based on their physico-chemical properties using machine learning techniques." *BMC bioinformatics* 22 (2021): 1-11.
33. Feng, Chang-Xue Jack, Zhi-Guang Samuel Yu, Unnati Kingi, and M. Pervaiz Baig. "Threefold vs. fivefold cross validation in one-hidden-layer and two-hidden-layer predictive neural network modeling of machining surface roughness data." *Journal of manufacturing systems* 24, no. 2 (2005): 93-107.