Machine Learning Regression Models for Predicting Hemolytic Concentration (HC_{50}) of Peptides Using Curated Activity Data

by

Ayushi Singh (MT23242)

Under the guidance of

Prof. G.P.S Raghava

Head and Professor

Submitted

in partial fulfillment of the requirements for the degree of

Master of Technology

to

Department of Computational Biology,

Indraprastha Institute of Information Technology Delhi

June, 2025

Certificate

This is to certify that the thesis titled" **Machine Learning Regression Models for Predicting Hemolytic Concentration (HC_{50}) of peptides using the Activity Curated Data"** being submitted by Ayushi Singh to the Indraprastha Institute of Technology Delhi, for the award of the Master of Technology, is an original research work carried out by her under my supervision. In my opinion the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

This thesis's results have not been submitted to any other university or institute for degree/diploma award in part or full.

June, 2025

Prof. G.P.S Raghava

Department of Computational Biology

Indraprastha Institute of Information Technology Delhi

New Delhi 110 02

Acknowledgement

I would like to sincerely express my gratitude and respect towards Prof. Gajendra P.S. Raghava from Indraprastha Institute of Information Technology, Delhi for being my supervisor and for exposing me to this wonderful topic of research and guiding me throughout. Besides my supervisor, I would like to thank PhD scholar Anand Singh Rathore, for his constant guidance, support, and motivation throughout the project. I would also like to thank the Department of Computational Biology and IT administrators at IIIT Delhi for providing me with all the necessary resources. Last but not least, I want to express my gratitude to my family and friends for their invaluable support and occasional encouragement during the writing of my thesis, which allowed me to do my study effectively and methodically.

Ayushi Singh
M.Tech CB (MT23242)

Table of Contents

List of abbreviations

HC_{50}	Hemolytic Concentration 50
RBC	Red Blood Cell
ML	Machine Learning
R	Pearson Correlation Coefficient
R^2	Coefficient of Determination
MAE	Mean Absolute Error
MSE	Mean Squared Error
SMILES	Simplified Molecular Input Line Entry System
REST API	Representational State Transfer – Application Programming Interface
APD	Antimicrobial Peptide Database
DAMPD	Database of Antimicrobial Activity and Structure of Peptides
UniProt	Universal Protein Resource
CAMP	Collection of Antimicrobial Peptides
MAP	Multiple Alignment Profile
SPSS	Solid-Phase Peptide Synthesis
AAC	Amino Acid Composition
DPC	Dipeptide Composition
PCP	Physicochemical Properties Composition
BTC	Bond Type Composition
ATC	Atom Type Composition
CTC	Conjoint Triad Composition
RRI	Residue Repeat Information
DDOR	Distance Distribution of Residues

SER	Shannon Entropy of Residues
SPI	Shannon Entropy of PCP Features
SPC	Split Composition
SEP	Split Entropy Profile
QSO	Quasi Sequence Order
XGBR	Extreme Gradient Boosting Regressor
RFR	Random Forest Regressor
ETR	Extra Trees Regressor
GBR	Gradient Boosting Regressor
SVR	Support Vector Regressor
KNNR	K-Nearest Neighbors Regressor
ADBR	AdaBoost Regressor
DTR	Decision Tree Regressor
DSSP	Define Secondary Structure of Proteins
PDB	Protein Data Bank
API	Application Programming Interface

List of Figures

List of Tables

Abstract

In this thesis, we first compiled hemolytic activity data of peptides in terms of hemolytic concentration (HC_{50}), defined as the concentration required to lyse 50% of red blood cells (RBCs). We then developed regression models using machine learning techniques to predict HC50 values, which serve as a key indicator of hemolytic potential. This activity data has been integrated into Hemolytik2 (http://webs.iiitd.edu.in/raghava/hemolytik2/), an updated and enhanced version of the Hemolytik database. Hemolytik2 is a manually curated and systematically organized resource that compiles experimentally validated hemolytic peptides from literature and public repositories, including the Antimicrobial Peptide Database (APD), UniProt, and the Dragon Antimicrobial Peptide Database (DAMPD). Over 5,000 of the 13,215 validated peptides in the database have known HC 50 values. Additionally, 2,569 peptides with experimentally established HC50 values against mammalian RBCs were used to train the regression models. With a Pearson correlation coefficient (R) of 0.660 and a coefficient of determination (R2) of 0.408, the top-performing model demonstrated a decent capacity for prediction. All things considered, Hemolytik2.0 is a useful platform for investigating the hemolytic characteristics of peptides and aids in the creation of computational tools meant to create safer and more efficient peptide-based drugs.

Chapter – 1

Introduction

Peptides have increasingly become central to modern drug discovery efforts owing to their remarkable pharmacological versatility, biological specificity, and relatively low systemic toxicity [1], [2]. As functional biomolecules, therapeutic peptides offer a unique intermediate between small-molecule drugs and large biologics, capable of targeting protein–protein interactions and other complex molecular mechanisms that were previously considered undruggable [3]. These features have rendered them highly successful in curing a wide range of diseases such as cancers, metabolic disorders, autoimmune diseases, and microbial infections [4]. The continued evolution of solid-phase peptide synthesis (SPPS), high-throughput screening, and chemical modification strategies has expanded the chemical diversity and stability of peptide-based drugs, driving their growth in clinical pipelines [5]. “Therapeutic peptides occupy a unique middle ground between small-molecule drugs and biologics,” [1].

Despite these developments, toxicity issues, notably hemolytic toxicity—the ability of peptides to compromise the membrane integrity of red blood cells (RBCs) and cause hemolysis—frequently impede the clinical development of peptide therapies [6], [7]. The release of haemoglobin into the plasma as a result of hemolysis can cause serious immunological reactions, oxidative stress, and kidney damage. Cationic, amphipathic, and membrane-active peptides often interact non-specifically with lipid bilayers, including RBCs, leading to membrane instability, lipid extraction, or pore formation [8]. Even while these interactions can occasionally be advantageous in antibacterial or anticancer settings, they pose a serious obstacle to systemic administration.

The peptide concentration dose at which, 50% of RBCs lyse under healthy conditions is defined by the HC_{50} (hemolytic concentration 50) experiment, which is frequently used to evaluate hemolytic potential [9]. Despite its value, this parameter is often neglected in the early stages of drug development since there

aren't enough trustworthy, complete, and centralized databases on hemolytic peptides. Predictive modelling and SAR research rely on structural and physicochemical characteristics, which are often missing from existing data, dispersed throughout publications, and inconsistently annotated. [10], [11].

The Hemolytik database was initially created to gather experimentally verified data on hemolytic and non-hemolytic peptides, enabling researchers to analyse peptide characteristics and identify trends related to hemolytic activity. However, its initial version lacked standards, computational integration, and structural annotations, limiting its usage.

Hemolytik 2.0 is an updated resource for large-scale predictive modelling, containing 13,215 peptide entries from reputable databases like UniProt, CAMP-R4, APD, and DAMPD. It offers comprehensive annotations including biological function, stereochemistry, origin, homolytic profile, terminal alterations, structural class, and sequence information, addressing issues in the original resource.

Hemolytik 2.0 is a comprehensive platform for studying hemolytic peptides, offering various bioinformatics techniques like SMILES search, BLAST search, and sequence alignment tools. It also integrates machine learning-based predictive toxicology for estimating new peptides' hemolytic potential. The platform's RESTful API allows seamless integration of high-throughput screening systems and external bioinformatics procedures. This dynamic environment is ideal for researchers to create safe and efficient peptide therapies, making it a valuable tool for research.

Hemolytik 2.0 is a peptide database that combines experimental data with computational tools, enabling a better understanding of structure-toxicity relationships, a crucial area often overlooked due to a lack of standardized data. It supports computational modelling, docking investigations, and SAR analysis [13].

The integration of validated hemolytic concentration values and machine learning-compatible features allows for the creation of regression models that accurately predict peptide toxicity levels, enabling the rational design and screening of analogues [14].

Hemolytik 2.0, a web interface and API, enhances accessibility and reproducibility for designing novel peptide therapeutics, academic training, hypothesis testing, and comparative studies across peptide datasets [15].

Hemolytik 2.0 is a crucial tool in peptide-based drug creation, providing a centralized, interactive repository for collaboration between experimentalists and computational biologists, thereby accelerating safer therapeutic discovery.

Chapter – 2

Compilation and Mining of Hemolytic Activity of Peptides for Dataset

INTRODUCTION

Peptides, comprising brief chains of amino acids, play vital roles in functions such as enzyme catalysis, hormone signaling, and immune responses [16]. They have drawn a lot of interest recently as possible therapeutic medications because to their remarkable accuracy in precisely targeting molecules linked to illness. They are also typically safe, simple to adjust, and the body tolerates them well. Advances in peptide design and screening methods have led to the discovery of several novel peptides with promising therapeutic applications for autoimmune diseases, metabolic disorders, cancer, and infections.

While peptides offer therapeutic promise, some pose toxicological risks, which must be assessed to ensure their safety in medical applications. Hemotoxicity is a severe kind of peptide drug toxicity in which the peptide breaks the membranes of red blood cells, causing damage [17]. For peptide-based therapies to be safe, hemolytic toxicity must be understood and predicted early in the drug development process.

The process by which red blood cells disintegrate and release haemoglobin into the circulation is known as hemolysis (Figure 1) [18]. This may occur when certain peptides harm red blood cells' outer membrane [19]. It can be caused by clumping together on the cell surface, weakening the membrane structure, or creating holes (pores) in the membrane [20]. Red blood cells may rupture as a result of these activities, which might have negative physiological effects. An evaluation metric known as HC_{50} is used by researchers to quantify a peptide's toxicity to red blood cells. The concentration of a peptide at which 50% of red blood cells are normally damaged is indicated by this value. Peptides with lower HC_{50} levels are more harmful to red blood cells and may cause excessive hemolysis, making early research and forecasting of hemolytic toxicity crucial in drug development for safety.

We introduce Hemolytik 2.0, an updated version of the original Hemolytik database, which provides improved data on peptide toxicity. It includes 13,215 peptide entries with about 8,000 distinct sequences, including amino acid sequence, biological source, known activity, cyclic or linear structure, and chemical modifications at the ends. This information allows researchers to easily identify and compare toxic and non-toxic peptides due to the information it offers on the peptide's hemolytic activity.

The revised version of Hemolytik 2.0 includes essential information like SMILES, chemical structure representation, and anticipated three-dimensional peptide forms.

A REST API has been implemented for easier access using automated tools. The database's peptide alignments in MAP format allow users to view sequences side by side and find conserved sections. The MAP format also identifies chemically altered sequences, such as terminal modifications, D-amino acids, or non-natural amino acids, making it easier for researchers to understand their impact on hemolytic potential and peptide activity.

[24].

Hemolytik 2.0 is a crucial tool for expediting the development of safer, more potent peptide-based medications by simplifying the identification and creation of non-hemolytic candidates.

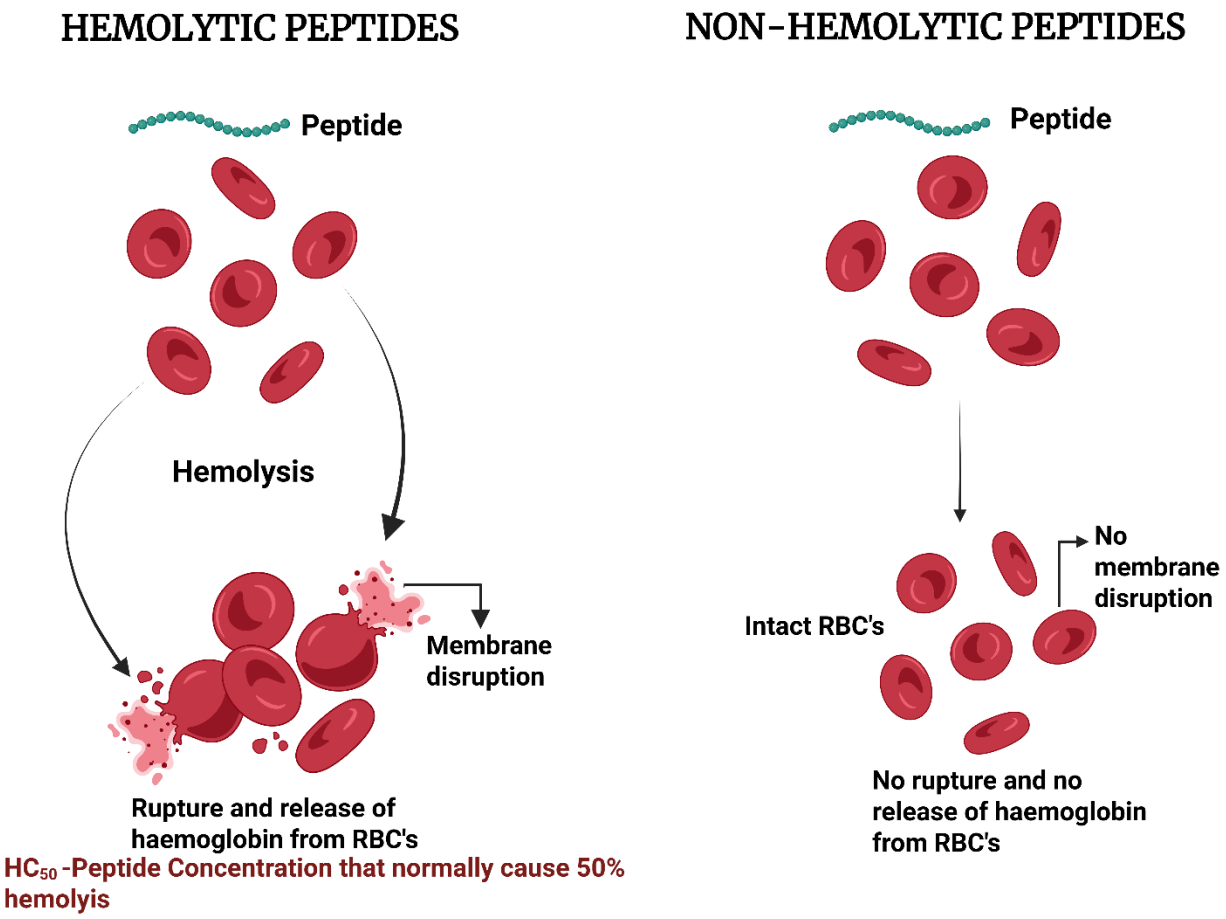


Figure 1: Illustration of how Red Blood cells are affected by Hemolytic vs. Non-Hemolytic Peptides

MATERIALS AND METHODS

Collection of Data

Hemolytik 2.0 is a comprehensive collection of scientifically data covering hemolytic and non-hemolytic peptides were retrieved through a targeted literature review via PubMed. Therefore, in order to find original research publications published between 2013 and 2024. 4,533 studies reporting experimental hemolysis tests were selected for manual data extraction process involved a query that included hemolysis, hemotoxin, peptide, and a review or systematic review, but did not include a review or systematic review."[25].
Along with literature mining, peptide entries were gathered from a number of reputable databases. These included APD3 (https://aps.unmc.edu) with 179 peptides [26], UniProt (https://www.uniprot.org) [27] with 268 peptides, DAMPD (http://apps.sanbi.ac.za/dampd/) [28] with 6 peptides, and CAMP R4 (http://www.camp.bicnirrh.res.in) [29] with 35 peptides. Hemolysis and "hemolytic" were among the keywords that were used to find pertinent content.

Web-based interfaces and database architecture

Following data collection and refinement, the development of the database infrastructure was undertaken, which was accomplished with MySQL and the Apache HTTP Server [30]. HTML, CSS, PHP, and JavaScript were used to develop the database's front-end interface [31]. The open-source and cross-platform interoperability of Apache, MySQL, and PHP led to their selection [32]. The scripts for the common gateway interface and database interface were all written in PHP and Perl [33]. As an object-related database management system (RDBMS), MySQL acts as the backend and supplies the commands required for data store and retrieval [34].

Data content of database

The Hemolytik 2.0 database comprises a total of 13,215 entries, each of them represents a peptide that is physiologically significant and has comprehensive annotations pertaining to its source, sequence, structural characteristics, and

hemolytic activity [35]. Every entry has a unique ID, and the PMID (PubMed ID) and the year of publication are used to link it to its source literature [36]. The sequence field, which includes each peptide's calculated length and amino acid sequence, is the central component of the dataset [37].

The MAP format in Hemolytik 2.0 database displays aligned peptide sequences, including chemically modified residues, to help visualize conserved regions and structural differences [38]. The database provides information on structural modifications, such as the presence of chemical groups at the N-terminal (nter) and C-terminal (cter) ends, as well as whether the peptide is cyclic or contains linkers (lyn_cyc), in order to capture the chemical complexity of peptides [39]. The stereochemistry of the amino acids is shown by the ldmix field, which shows whether the sequence comprises L, D, or a blend of the two [40]. The non_nat column helps users differentiate between naturally occurring and chemically modified peptides by indicating the presence of non-standard or synthetic residues [41].

Apart from structural characteristics, the database records crucial biological information. While nature indicates whether the peptide is synthetic or natural [42], the name field gives each peptide a reference name or identification [43]. With an emphasis on the peptide's hemolytic potential, the activity field explains its known or observed biological role [44]. Additionally, the source and origin columns show the peptide's biological origin and method of derivation, respectively [45]. Peptides that have been experimentally confirmed to be non-hemolytic are marked in the non_hem column, while information on experimental structural data is noted under the exp_str column [46].

Hemolytik 2.0 database is a valuable resource for researchers studying therapeutic peptides, peptide-based toxicity, and safer peptide drug design [47].

ORGANISATION OF THE DATABASE

The tools of the database are categorized as:

1) **Search**

   i) Basic Search – This search tool version focuses on basic functionality to retrieve data from the database. It enables users to search the database using keywords in fields, such as PMID, sequence, nature, length, MAP format, stereochemistry, source, hemolytic activity, origin, C-terminal and N-terminal modifications, and more [48].

   ii) Advanced Search –In this tool, the user can fetch data from the database by adding several fields and entries at a time [49].

   iii) Peptide Search – The user has two choices when using this tool:
   (a) **Exact peptide** – The user can retrieve data for identical peptides in the database.

   (b) **Containing peptide** – this feature gives information based on the peptide query that the user enters in the box [49].

   iv) SMILES Search – This tool allows the user to search a particular SMILES notation in SMILES format against the Hemolytik 2.0 database. Sequences were converted to SMILES format using the RDKit tool [50].

2) **Browse** – The purpose of this browsing tool is to make it easier for the user to find data in the Hemolytik 2.0 database in a structured way. Here, users may explore a variety of areas such as the peptides' source (human, pig, horse, etc.), type (linear, cyclic), stereochemistry (L, D, or mix), and non-hemolytic nature using a user-friendly interface. The user may see the count of each field in addition to accessing the PMID, C-ter, N-ter, and other modifications. Furthermore, each field is connected to the database to provide information about that particular field, making it more informative [51].

3) **Analysis Tools**

i) BLAST – With this tool, users may submit FASTA format peptide sequences and modify fields like weight matrix and expectation value to conduct similarity-based searches against hemolytic and non-hemolytic peptides [52].

ii) Smith-Waterman – For short peptides, this tool performs a similarity search more successfully. Users can enter several sequences in FASTA format to carry out this analysis [52].

iii) Mapping – The platform enables alignment of hemolytic peptide segments with longer protein or peptide sequences through similarity analysis. Based on a protein sequence, it extracts related peptides, compares input peptides to every peptide in the Hemolytik 2.0 database, and allows users to submit sequences to identify sections that are identical [53].

iv) Structure Alignment – The user aligns the structure of a PDB file with the peptide structure whose ID is provided in the box after entering the file in the given box [53].

4) **REST API** – A RESTful API has been added to Hemolytik 2.0 to enable automated data access. This enables programmed data retrieval by users based on parameters such as the source organism, sequence classification, or known hemolytic potential. Responses from the API are provided in JSON format, allowing for easy integration with a range of bioinformatics tools and workflows. Additionally, users can filter and interpret the data to meet their own needs [50].

5) **Download** – The entire database from Hemolytik 2.0 is downloadable. Also, PDB files with predicted tertiary structures are provided. Every open-access reference article from PubMed that was utilized to compile this resource can also be downloaded [51].

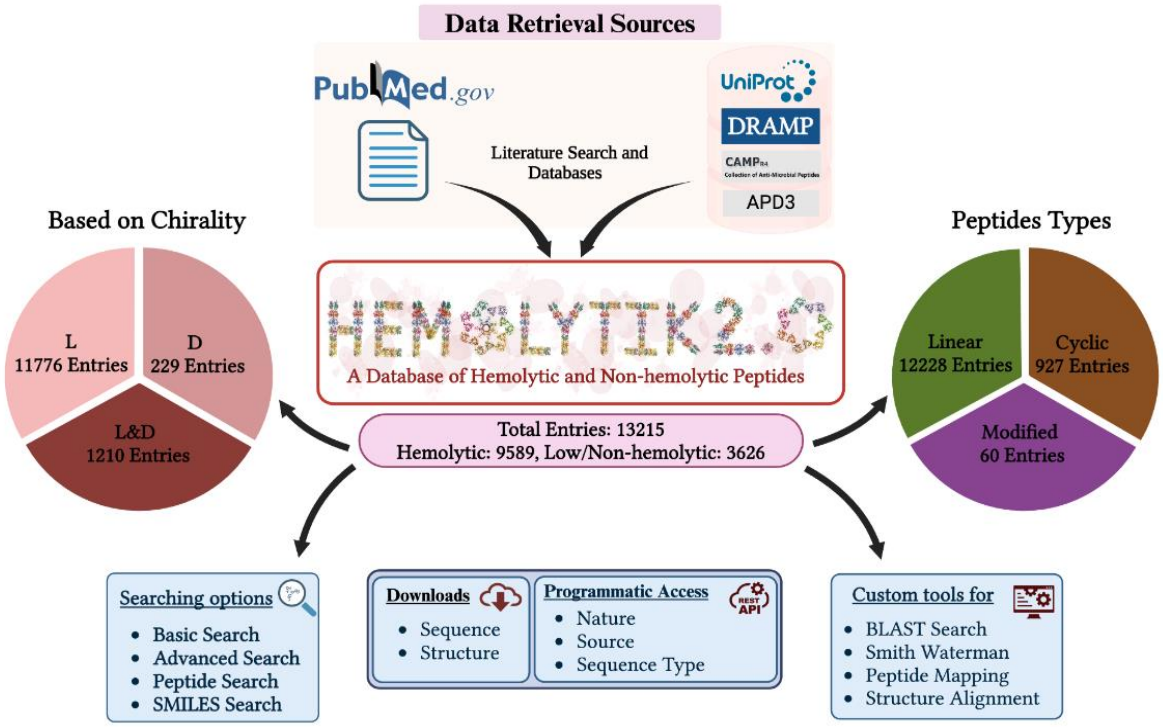


Figure 2: Hemolytik 2.0 Architecture

DATABASE STATISTICS

Hemolytik 2.0, the revised edition of Hemolytik database contains 13215 total entries that were taken from publications, websites and other databases. The 13215 entries include 9589 hemolytic and 3626 non-hemolytic peptides [44]. The statistics we had is divided into two categories: -

1. <u>All Hemolytic Peptides Statistics</u>

   All hemolytic peptides statistics are included in this section. Assays on Rbc's of various species were used to measure hemolytic activity. We have discovered several sources such as human, horse, pig, sheep, goat, rats, rabbit, fish, etc., therefore, we combined the data we obtained 20 sources, as

fig illustrates. The most often discovered source human (9255), followed by horse (1062), sheep (840), and so on. Their entries included various doses like human at 50 µM and 100 µM.

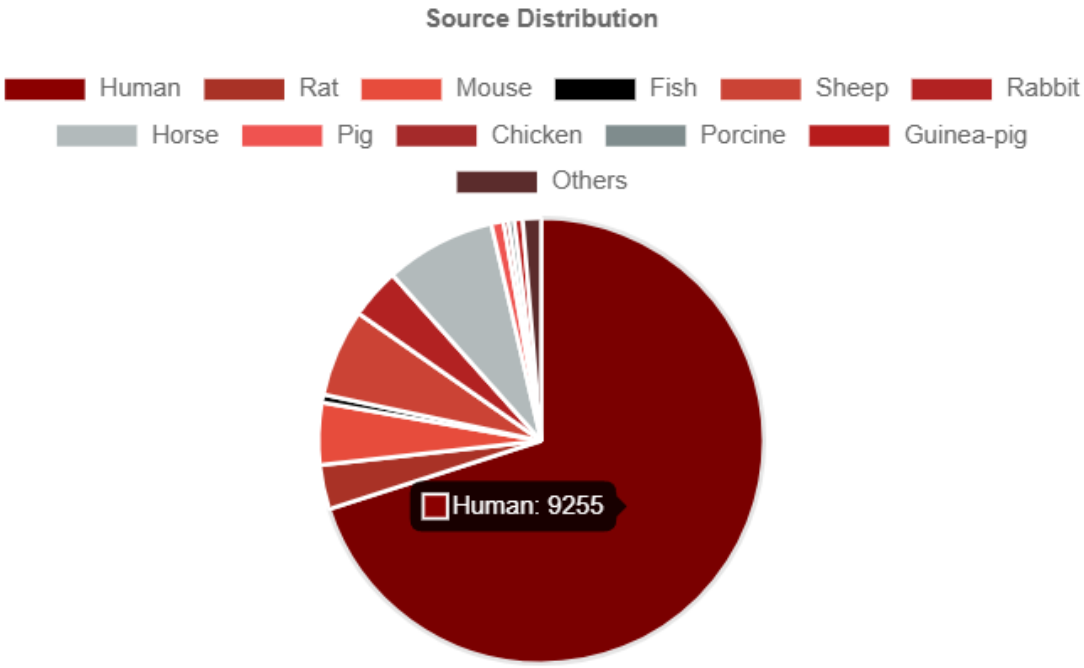


Figure 3: Red blood cells sources used to measure hemolytic activity

Different therapeutic categories are represented by the peptides that are a part of Hemolytik 2.0 database. The majority of the entries have an antimicrobial character (5760), followed by an antibacterial nature (2841), reflecting the various biological activities that the peptide represents, including antimicrobial, anticancer, antibacterial, cell-penetrating, etc. [54]. The peptides have been further classified according to their shape (Linear, Cyclic) and stereo-chemistry (L, D, or mixed amino acids). In all, there are 12211 entries for linear peptides [55], 906 entries for cyclic [55], 11728 entries for L, 229, and 1210 forms of amino acids, respectively [55]. Fig. shows the statistics. Additionally, we have collected data on peptides that

have experienced chemical changes, such as those to the C-terminus and N-terminus.

Leucinol (Lol, -CH(NH_2)-CH_2OH), Cha=β-cyclohexylalanine, and modified terminals (2-(6-Methoxy-2-Naphthyl) propionic acid) are examples of amino acid chemical modifications. The database has 1935 entries describing hemolytic peptides[56] with these kinds of chemical modifications.

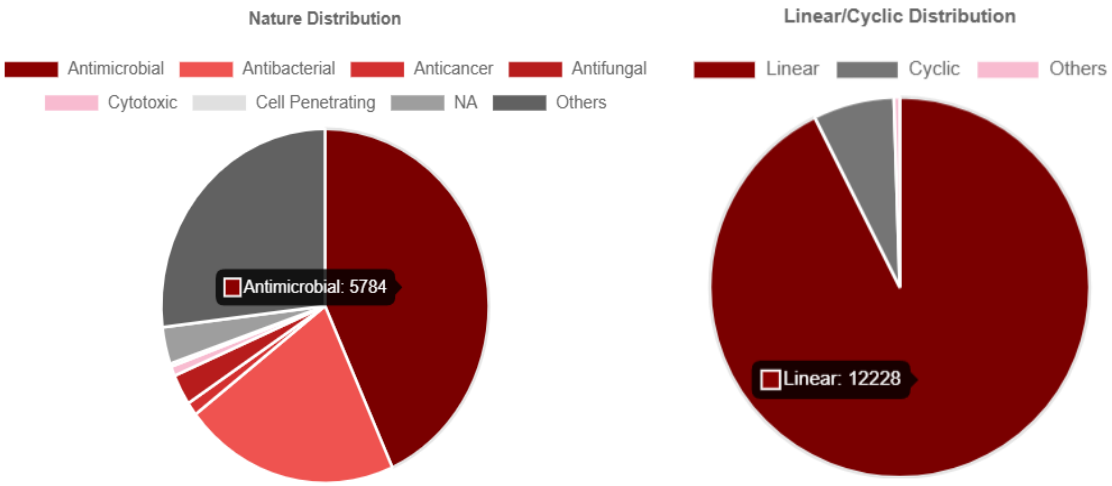


Figure 4: Biological function statistics (A) and linear/cyclic distribution of peptides (B)

2. Natural Hemolytic Peptides Statistics

   In this part of stats, we have all same categories such as sources of RBCs, biological function, stereochemistry and chirality, etc. But here we have 7654 entries in the database describing naturally occurring hemolytic peptides [44], [45].

COMPARISON WITH THE PREVIOUS VERSION

The Hemolytik database, first launched in 2012, was developed to provide a curated collection of experimentally verified peptides that are hemolytic and non-

hemolytic [12]. It included about 2,970 entries, most of which included sequence details, the natural or synthetic origin of the peptide, and its hemolytic activity. The database lacked structural data, comprehensive annotations, sophisticated search capabilities, and programmatic access, while being one of the earliest resources devoted to peptide-induced red blood cell lysis [57].

Hemolytik 2.0, a significant upgrade, was published in 2024 to fill in these deficiencies [44]. By adding chemically changed peptides containing D-amino acids, cyclic structures, and terminal modifications, it greatly increased the database's size to 13,215 items [46]. The addition of secondary structure annotations using DSSP and anticipated tertiary structures using PEPstr, which allow for structural and functional analysis, is one of the major developments [58].

In order to provide deeper insights and better data linkage, Hemolytik 2.0 additionally connects each item to other databases such as Swiss-Prot, PDB, and PubMed [59]. Hemolytic peptides count increased from 1750 to 5598 [44] and non-hemolytic peptides from 295 to 2410 [44]. Also, the number of RBCs sources used in assays has grown from 17 to 20 [60].

Now, the platform has a mobile-friendly UI, advanced keyword search, and—most importantly—a RESTful API that enables programmatic data retrieval for pipeline or tool integration [31]. All of these improvements combine to create Hemolytik 2.0 a far more reliable and intuitive platform for researching peptide toxicity and creating safer therapeutic peptides [44], [60].

Table 1- Comparison of Hemolytik 2.0 with the previous version

Comparison Criteria	Hemolytik (2012)	Hemolytik 2.0 (2024)
Total Entries	2970	13215
Total Hemolytic Peptides	2651	9589
Total Non-Hemolytic Peptides	319	3626
Unique Hemolytic Peptides	1750	5498
Unique Non-Hemolytic Peptides	295	2410
Sources of RBC's	17	20
Chemical Modifications	221	2548
Structural Info	Not available	Predicted tertiary structures (PEPstr), secondary structure (DSSP)
External Links	Limited	Linked to UniProt, IEDB, PDB, PubMed, TrEMBL
Search tool	Basic keyword search	Advanced search with filtering
REST API	Not available	Available for programmatic access
Mobile Compatibility	Desktop only	Mobile-friendly design

DISCUSSION

The Hemolytik 2.0 database includes more than 13,000 peptide entries with comprehensive details on their origins, structures, sequences, and hemolytic

activity. It is particularly useful for developing novel peptide-based medications as it focuses on assisting researchers in determining which peptides are safe and which might harm red blood cells (hemolysis).

The majority of the database's entries are hemolytic peptides, which reflects the increased interest in researching their possible use as antibacterial or anticancer drugs. Important information is included in each entry, including the length of the peptide, any chemical alterations, whether it is natural or synthetic, and the kind of amino acids utilized (L, D, or mixed). Additionally, it uses PubMed IDs to connect each peptide to its original research.

The database has adequate information for researchers to examine structure–function correlations and create machine learning models to forecast hemolytic behaviour, even if it lacks receptor data and 3D structural models for all peptides. There may be missing data in certain entries, however coverage will be improved in next updates.

All things considered, Hemolytik 2.0 is a useful tool for fundamental studies as well as real-world uses in toxicity prediction and drug development.

CHAPTER-3

Machine Learning Regression Models for Predicting HC_{50}

INTRODUCTION

Peptides, short amino acid sequences, are essential for various several biological functions and have gained popularity for the creation of novel drugs in recent years due to their ease of design, general safety, and selectivity [1], [3], [61]. Peptide therapies, which treat infections, cancers, metabolic disorders, offer benefits over conventional small-molecule drugs, including improved target selectivity and fewer adverse effects [2], [3], [61].

Peptides can cause hemolysis, a process by where membrane of red blood cells (RBCs) ruptures and haemoglobin is released into the circulation [18], [19]. This can lead to serious health issues like anaemia, organ damage, or even death if the toxin level is too high [17], [62]. Therefore, it's crucial to evaluate the hemolytic potential of peptides while developing new drugs [6], [7].

Researchers use a metric known as HC_{50} (Hemolytic Concentration 50) to measure the harmfulness of a peptide to red blood cells. The peptide concentration at which half of the red blood cells in a sample are destroyed is denoted as HC_{50}. A high HC_{50} indicates a safer peptide, a low HC_{50} value indicates severe toxicity [20], [21]. Laboratory studies, which can be expensive and time-consuming are necessary to determine HC_{50} experimentally [9], [22].

Machine learning (ML) can assist in this situation. Machine learning algorithms may be trained to predict HC_{50} values using patterns in peptide sequences [14], [33], [63]. Also, "Machine-learning-based modelling enables rapid hemolytic-toxicity prediction," observe Zhao et al. [14]. Scientists can save time and money by estimating the hemolytic concentration in silico (using a computer) using trained machine learning models rather than conducting lab tests for each novel peptide [25], [33].

In this chapter, we focused on developing regression models, a kind of machine learning method for forecasting continuous values such as HC_{50}. We make use of an empirically validated peptide dataset as well as a number of features extracted from the amino acid sequences of the peptides [15]. The physicochemical and structural characteristics of the peptides are described in part by these features [6], [64].

We assess several regression methods, compare their results using statistical measures (e.g., R^2 score, Mean Absolute Error, etc.), and determine which model is most effective in predicting HC_{50} values [14], [63]. The findings of this study have might lessen the necessity to extensive laboratory testing, speed up the creation of safer peptide-based drugs, and aid in the early screening of peptide toxicity [33], [63].

MATERIALS AND METHODS

Dataset compilation and preprocessing

We collected a total of 2,569 peptide sequences by combining data from two sources: the HemoPI2 dataset and the Hemolytik2 database. From HemoPI2, we obtained 1,926 sequences [65], and from the Hemolytik2 dataset of 13,215 peptides, we selected 643 unique sequences that were not already present in HemoPI2 [12], [44]. During this process, we removed duplicate sequences as well as those that were too long or too short to maintain uniformity [66]. Outliers were also eliminated according to length distribution, as shown in the fig.

After filtering, we finalized a dataset of 2,569 unique peptides. This dataset was then split into two parts: 80% for cross-validation and the remaining 20% for independent testing [67]. The independent set was saved as the evaluation dataset for model validation and final performance analysis.

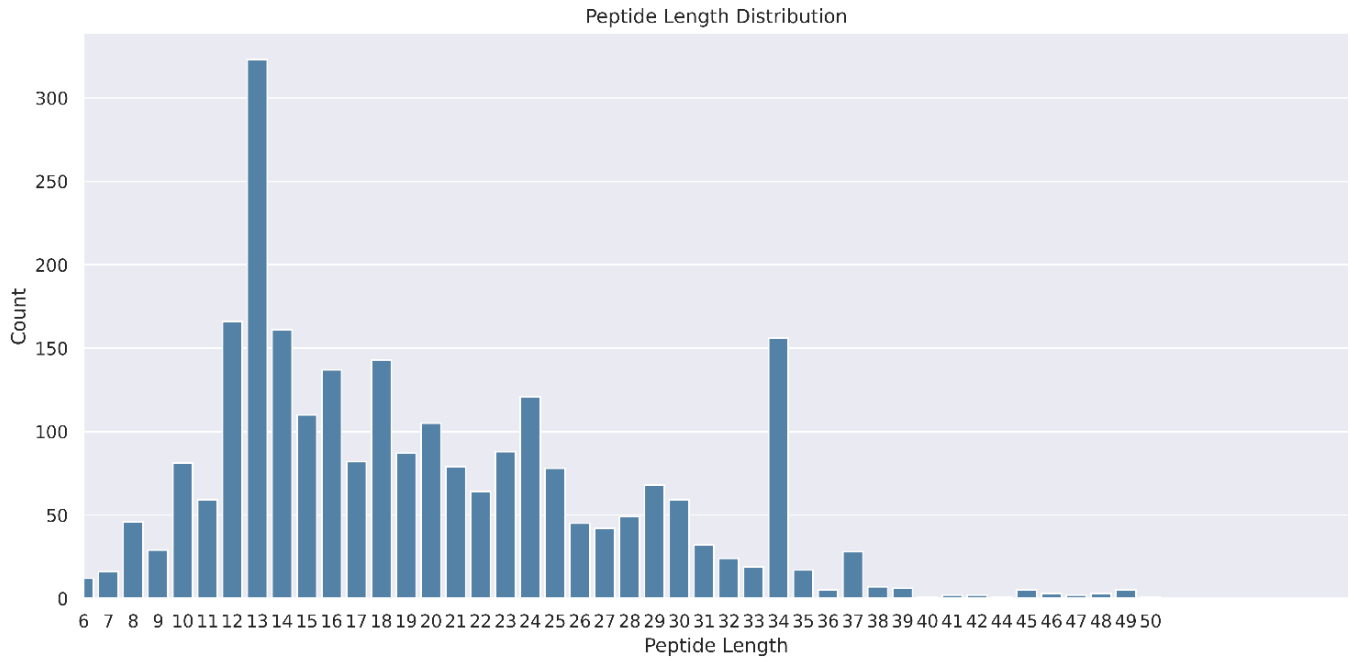


Figure 5: Length distribution of the peptides

Feature Extraction

We have generated over 8900 features using standalone method called Pfeature [64f]. In total, we calculated 8974 features which were used by models for predictions. The details and length of each feature is listed in table.

Feature Extraction using Pfeature

In this work, we have utilized a large collection of features that were retrieved using the Pfeature64 program to numerically represent peptide sequences for hemolytic activity prediction based on machine learning. AAC (20 features), DPC (400 features), TPC (800 features) [64]. To encode crucial biological characteristics including charge, polarity, and hydrophobicity, the Physicochemical Properties Composition (PCP, 30 features) was utilized [68]. Bond Type Composition (BTC, 4 characteristics) and Atom Type Composition (ATC, 5 features) take into consideration the bond types and elemental composition of the peptides [69]. Conjoint Triad Composition (CTC, 343 characteristics) took into

account the frequency of three successive amino acids arranged according to the polarity and side-chain volume [70]. To quantify sequence complexity and variety, entropy-based descriptors were employed, such as Shannon Entropy of Residues (SER, 20 features) and Shannon Entropy of PCP features (SPI, 26 features) [71]. Repetitive patterns and the spatial distribution of amino acids were captured by include Residue Repeat Information (RRI, 20 features) and Distance Distribution of Residues (DDOR, 20 features) [72]. Regional amino acid composition and entropy were recorded using Split Composition (SPC) and Split Entropy Profile (SEP) across the N-ter, middle, and C-ter regions of the sequence. Quasi-Sequence Order (QSO, 42 characteristics) quantified the link between residues based on physicochemical distance, therefore capturing structural motifs and sequence-order correlations [73]. Furthermore, as scalar descriptors, peptide length and molecular weight were included. With 1192 dimensions per peptide, this feature set served as the foundation for Hemolytik 2.0's machine learning-based predictive modelling.

Table 2- A list of all the calculated features and its vector length

Name of the Feature	Feature vector length
Amino acid composition (AAC)	20
Amphiphilic pseudo amino acid composition (APAAC)	23
Atom composition (ATC)	5
Bond composition (BTC)	4
Conjoint Triad Calculation (CTC)	343
Dipeptide composition (DPC)	400
Distance distribution of residue (DDOR)	20
Physicochemical Properties Composition (PCP)	30
Pseudo amino acid composition (PAAC)	21
Quasi-sequence order (QSO)	42
Residue Repeat Information (RRI)	20
Shannon Entropy of Physicochemical Property (SPC)	25
Shannon Entropy of Residues (SER)	20
Shannon-Entropy of Protein (SEP)	1
Tripeptide composition (TPC)	8000

Correlational Analysis

Correlation analysis shows how peptide features are related to their drug activity and safety. It highlights important determinants and possible trends by quantifying the direction and degree of correlations between variables [74]. The table 3 displays the top features of AAC, DPC, and physico-chemical characteristics that show a correlation with the hemolytic peptides' HC_{50} concentration. Additionally, there were significant relationships between PCP characteristics such hydrophobicity, charge, and polarity and HC_{50} [68][75].

Table 3- Top AAC, DPC and PCP features correlated with HC_{50} value of hemolytic peptides.

Single Amino Acid Composition	Correlation	Di-peptide Composition	Correlation	Physico-Chemical Features	Correlation
Lysine AAC_K	0.229	Lysine-Lysine DPC1_KK	0.150	Hydrophobicity PCP_HL	0.233
Valine AAC_V	0.120	Isoleucine-Lysine-DPC1_IK	0.138	Polarizability PCP_PC	0.226
Methionine AAC_M	0.062	Methionine-Alanine DPC1_MA	0.130	Bulkiness PCP_BS	0.226
Histidine AAC_H	0.054	Lysine-Glycine-DPC1_KG	0.118	Z-scale descriptor Z4 PCP_Z4	0.180
Leucine AAC_L	0.046	Alanine–Lysine DPC1_AK	0.112	Secondary Structure – Helix- PCP_SS_HE	0.147
Glutamic Acid AAC_E	-0.368	Isoleucine–Glutamic Acid-DPC1_IE	-0.435	Net Charge PCP_NC	-0.356
Glycine AAC_G	-0.191	Tyrosine–Glycine DPC1_YG	-0.414	Acidic Residue Content--PCP_AC	-0.356
Tyrosine AAC_Y	-0.163	Glutamic Acid–Glycine-DPC1_EG	-0.407	Number of Turns in Secondary Structure-PCP_NT	-0.250
Aspartic Acid AAC_D	-0.154	Aspartic Acid–Glycine-DPC1_DG	-0.381	Z-scale descriptor Z3-PCP_Z3	-0.193
Cysteine AAC_C	-0.111	Glutamic Acid–Asparagine-DPC1_EN	-0.358	Secondary Structure – Coil-PCP_SS_CO	-0.166

Stronger hemolytic activity is suggested by a positive correlation, which implies that greater feature values result in higher HC_{50}. Higher feature values result in lower HC_{50}, indicating stronger hemolytic activity, when there is a negative connection. These findings aid in our comprehension of the characteristics of peptides that influence their hemolytic action. Figure 6 displays the top 10 positive and negative correlated features.

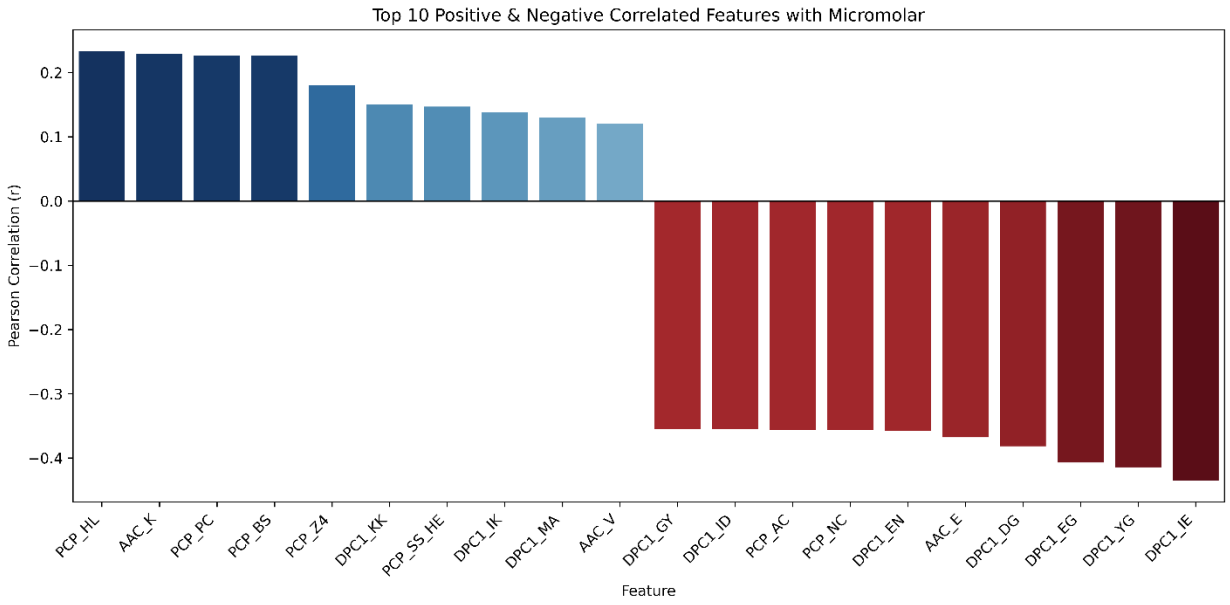


Figure 6: Top 10 positively and negatively correlated features with HC50 concentration, highlighting key AAC, DPC, and PCP attributes influencing hemolytic activity.

Predictive target for Regression Analysis

The pHC_{50}, or negative logarithm of the HC_{50} value, was the objective of our regression model in this study. It is computed using the following formula:

$$\mathbf{pHC_{50} = \log_{10}(HC_{50})}$$

Thus, we used a logarithmic scale to the HC_{50} values. By doing this, the model's performance may be improved and the large range of variables is managed [76]. Drug research and bioinformatics frequently employ this kind of modification to manage activity data [77]. It makes it easier to compare the hemolytic activity of various peptides and increases prediction accuracy [78].

Cross-Validation Approach

We used accepted bioinformatics procedures to guarantee the robustness and dependability of our models. A random division of the dataset was made, with 20% serving as an independent test set and the remaining 80% for cross-validation. The independent set was kept entirely apart and wasn't utilized for hyperparameter tuning, testing, or model training [79].
The 80% cross-validation set employed a five-fold cross validation technique [80]. Data had to be divided into five equal parts for this. The model was trained in four parts and tested in the remaining part of each round, make each part the test set once, this process was carried out five times. Using this method, we were able to consistently assess the model's performance across several data splits [81]. After the models were completed, their performance was evaluated using the independent (unseen) test set. It is essential to compare outcomes on this held-out group in order to verify any model's actual predictive ability [82].

REGRESSION RESULTS

Regression analysis, which is frequently quantified using the half-maximal hemolytic concentration (HC_{50}), was employed in this study to predict the hemolytic activity of peptides quantitively [83]. To standardize the data, we used a log transformation because HC_{50} values vary greatly and frequently exhibit skewed distributions [84]. “Log-transforming bioactivity data often improves regression stability,” Wang and colleagues note [84]. In order to improve regression model performance, this transformation stabilizes variance and creates a more linear connection between features and the output variable [85]. Using these log-transformed HC_{50} values as the target variable, all models were trained.

The study used various machine learning regressors, including tree-based ensemble models such as XGBR, RFR, ETR, and GBR, were used to create prediction

models [86]. Other models we evaluated were LR, KNNR, SVR, ADBR, and DTR [87]. To ensure consistency, these models were trained using feature sets previously extracted including individual descriptors and their combinations.

This study used a dataset of 2,569 peptides with experimentally obtained HC_{50} values to build the regression models. A two-step process was used to ensure robustness and avoid overfitting. Models were first trained and then validated on the training dataset using k-fold cross-validation [88]. A separate test set was set aside for the final analysis to provide an objective assessment of model performance.

Performance metrics

The model's performance was evaluated using standard regression measures, such as Pearson correlation coefficient (R), coefficient of determination (R^2), MAE, MSE [89]. Together, these measures reveal how well each model predicts HC_{50} values: lower MAE and MSE suggest less prediction error, while higher R and R^2 values show better correlation and variance explanation.

In every case, tree-based models performed better than others, and the best regressor was XGBR. In this study, with a R of 0.661, R2 of 0.408, MAE of 0.593, and MSE of 0.693 on the independent test set, the XGBR model trained on the ALLCOMP feature set produced the best results. These results validate the model's dependability by showing a low error and a significant predictive correlation. The DPC+PCP feature set with RFR and the BTC+CTC feature set with RFR were close competitors, but their performance metrics were marginally worse.

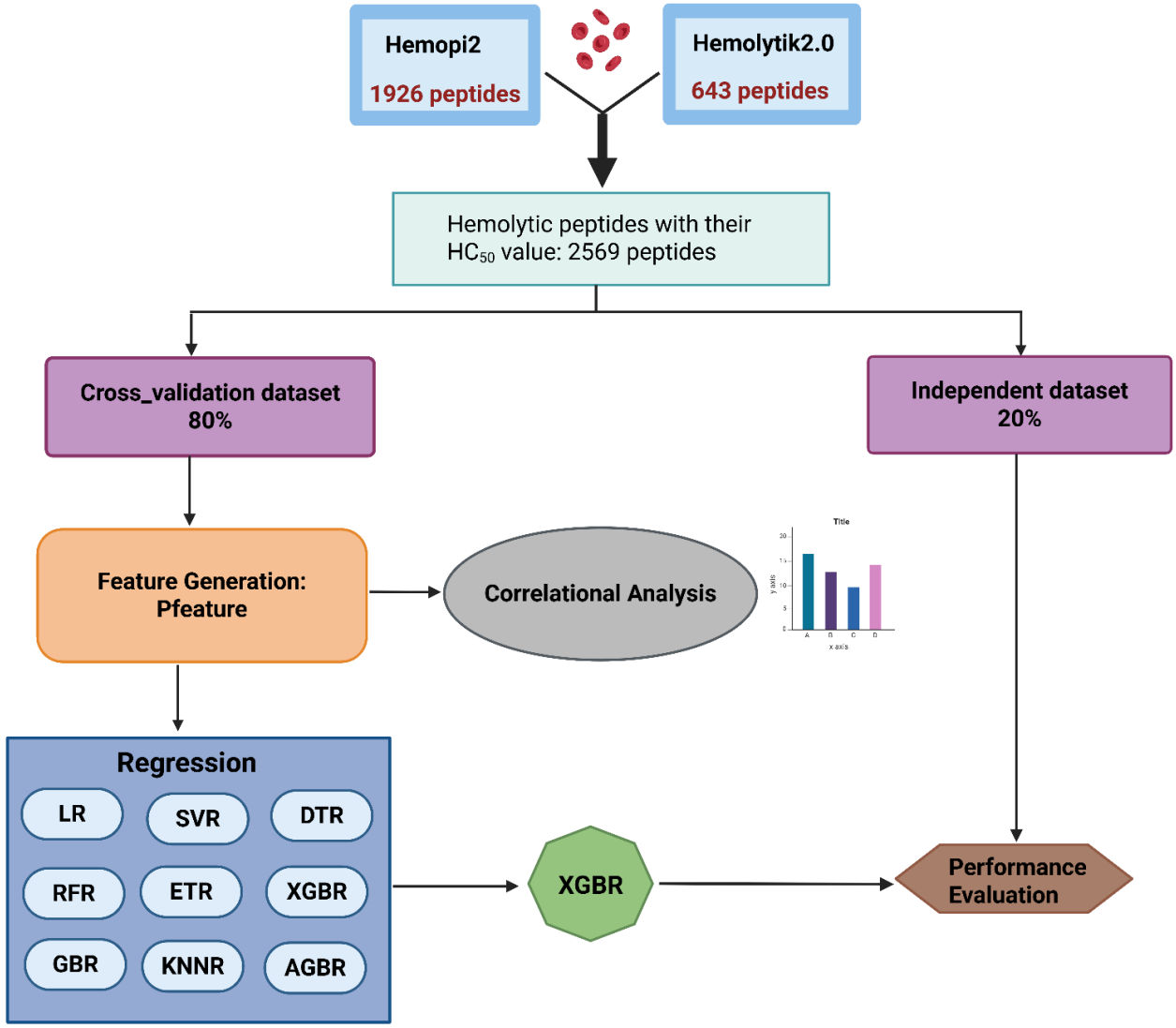


Figure 7: Schematic workflow of model for prediction of hemolytic peptides

In this study, a fair and relevant evaluation was ensured by choosing the best models based on a combination of strong correlation (R, R^2) and low error (MAE, MSE) [90]. Overall, the findings demonstrate the effectiveness of ensemble-based regression techniques in predicting hemolytic activity, particularly when combined with comprehensive peptide features [91].

Evaluation of Regression Models for HC_{50} Prediction

The performance of various regression models applied to different peptide feature sets for predicting log-transformed HC_{50} values is shown in Table 4. The

evaluation was conducted using four common metrics: Pearson correlation coefficient (R), mean absolute error (MAE), mean squared error (MSE), and coefficient of determination (R^2). With the lowest MAE (0.593) and MSE (0.693), the greatest R (0.661) and R^2 (0.408), and the best overall performance of all the models, XGBR utilizing the ALLCOMP feature set demonstrated little prediction error and strong correlation [92].

Table 4- Performance metrics of ML regressor model are evaluated using different features derived from Pfeature on independent dataset based on HC_{50} value

Features	Model	R	R^2	MAE	MSE
AAC	SVR	0.552	0.264	0.607	0.879
DPC	ETR	0.607	0.344	0.616	0.783
PCP	RFR	0.612	0.334	0.611	0.795
AAC+DPC	RFR	0.612	0.334	0.611	0.795
DPC+PCP	RFR	0.651	0.398	0.588	0.719
AAC+PCP	ETR	0.567	0.275	0.632	0.866
AAC+DPC+PCP	ETR	0.634	0.383	0.586	0.736
ATC+BTC	ETR	0.634	0.383	0.586	0.736
ATC+CTC	ETR	0.619	0.357	0.614	0.759
BTC+CTC	RFR	0.638	0.386	0.605	0.725
PAAC+APAAC	ETR	0.55	0.255	0.612	0.872
SER+SPC+SEP	GBR	0.608	0.33	0.672	0.793
APAAC+TPC	ETR	0.559	0.266	0.612	0.864
RRI+DDOR	XGBR	0.564	0.278	0.637	0.862
QSO+APAAC	SVR	0.544	0.252	0.596	0.88
ALLCOMP	**XGBR**	**0.661**	**0.408**	**0.593**	**0.693**

Top Three Rankings for Regression Models

The top three combinations of model and feature are shown in the table below. BTC+CTC with RFR (R = 0.638, MSE = 0.725) came in second, followed by DPC+PCP with RFR (R = 0.651, MAE = 0.588). These findings show how well ensemble models like as XGBR and RFR perform when combined with useful peptide features.

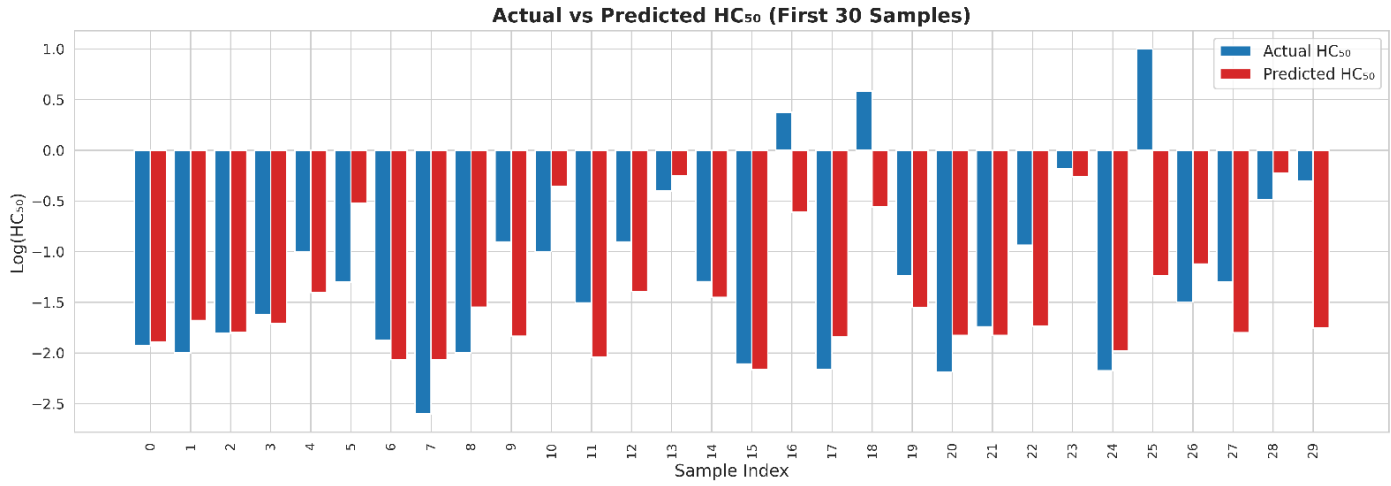
Table 5- HC_{50} prediction-based performance comparison of the top three regression models

Rank	Features	Model	Dataset	R	R^2	MAE	MSE
1.	**ALLCOMP**	**XGBR**	**Cross_valid**	**0.714**	**0.488**	**0.388**	**0.277**
			Independent	**0.661**	**0.408**	**0.593**	**0.693**
2.	DCP+PCP	RFR	Cross_valid	0.662	0.438	0.369	0.252
			Independent	0.651	0.398	0.588	0.719
3.	BTC+CTC	RFR	Cross_valid	0.655	0.424	0.387	0.265
			Independent	0.638	0.386	0.605	0.725

Table 6- Comparison of Actual vs. Predicted pHC_{50} Values for Peptide Samples

Actual HC_{50}	Predicted HC_{50}
-1.92942	-1.88993
-2	-1.67976
-1.80618	-1.79814
-1.62201	-1.71264
-1	-1.4038
-1.30103	-0.52324
-1.87506	-2.06681
-2.60206	-2.06801
-2	-1.54841
-0.90309	-1.8359
-1	-0.35769
-1.50515	-2.0414
-0.90309	-1.39334
-0.4014	-0.2511
-1.29469	-1.45171

Table 7- The model's predictive accuracy is demonstrated through a bar plot comparing actual and predicted pHC_{50} values for individual peptide samples.



Chapter-4

CONCLUSION

Peptides are gaining popularity as viable alternatives for creating novel medications. This is because they are incredibly effective, quickly metabolized by the body, and have a high accuracy in targeting particular disorders. These characteristics make them appropriate for treating a wide range of illnesses, such as autoimmune disorders, cancer, and infections. They have really already demonstrated promise in various fields, including the battle against viruses, immune system stimulation, and the destruction of cancerous cells and dangerous microorganisms.

However, hemotoxicity—particularly the hemolytic activity, or capacity to rupture red blood cells—is a major obstacle to the use of peptides as medications. This may render advantageous peptides unfit for human consumption.

Hemolytik 2.0 is an updated version of the original Hemolytik database containing data on over 13,215 peptides experimentally tested. It enhances research on peptide-based drugs, focuses on safety and hemolytic activity.

Hemolytik 2.0 provides data collection and tools for identifying toxic peptides, enhancing the safety and success of drug candidates by removing or altering them early in the design process.

Hemolytik 2.0 a computational biology tool, offers improved design, improved data, and new features like hemolytic activity strength prediction, making it valuable for scientists, researchers, and drug.

BIBLIOGRAPHY

[1] A. Lau and P. Dunn, “Therapeutic peptides: Historical perspectives, current development trends, and future directions,” *BioDrugs*, vol. 35, no. 3, pp. 245–264, 2021.

[2] D. Fosgerau and T. Hoffmann, “Peptide therapeutics: current status and future directions,” *Drug Discovery Today*, vol. 20, no. 1, pp. 122–128, 2015.

[3] G. Muttenthaler et al., “Trends in peptide drug discovery,” *Nature Reviews Drug Discovery*, vol. 20, no. 4, pp. 309–325, 2021.

[4] S. H. Lau and M. H. Dunn, “Applications of peptide therapeutics in treating cancer and infectious diseases,” *Frontiers in Pharmacology*, vol. 12, pp. 675–689, 2021.

[5] M. Goodwin and B. Simerska, “Advances in solid-phase peptide synthesis and purification,” *Peptide Science*, vol. 110, no. 4, e24156, 2018.

[6] H. Jiang et al., “Membrane-disrupting peptides: New generation of antimicrobial agents,” *Acta Pharmaceutica Sinica B*, vol. 11, no. 2, pp. 329–345, 2021.

[7] N. Porto et al., “Toxicity and hemolytic activity of antimicrobial peptides: A systematic review,” *Frontiers in Microbiology*, vol. 13, pp. 892132, 2022.

[8] K. Wimley, “Describing the mechanism of antimicrobial peptide action with the interfacial activity model,” *ACS Chemical Biology*, vol. 5, no. 10, pp. 905–917, 2010.

[9] M. Wang et al., “Assessment of hemolytic toxicity of antimicrobial peptides using red blood cells,” *Journal of Visualized Experiments*, no. 153, e60125, 2019.

[10] S. Agrawal et al., “Structure–activity relationship studies of hemolytic peptides: Challenges and opportunities,” *Molecular Informatics*, vol. 39, no. 5, e1900041, 2020.

[11] T. Raghava and S. Gautam, “Database resources for antimicrobial and hemolytic peptides,” *Briefings in Bioinformatics*, vol. 23, no. 1, bbab327, 2022.

[12] R. Chaudhary et al., “Hemolytik: A database of experimentally determined hemolytic and non-hemolytic peptides,” *Nucleic Acids Research*, vol. 44, D1, pp. D489–D493, 2016.

[13] S. Kang et al., “Peptide toxicity prediction: Past, present, and future directions,” *Briefings in Bioinformatics*, vol. 23, no. 6, bbac437, 2022.

[14] Y. Wu et al., “Machine learning approaches for identifying and designing hemolytic peptides,” *Bioinformatics Advances*, vol. 2, no. 1, pp. 1–10, 2022.

[15] G. B. Patel and M. N. MacLean, “Computational databases for peptide drug design: A review of tools and platforms,” *Drug Discovery Today*, vol. 28, no. 2, pp. 415–423, 2023.

[16] D. J. Craik, D. P. Fairlie, S. Liras, and D. Price, “The future of peptide-based drugs,” *Chemical Biology & Drug Design*, vol. 81, no. 1, pp. 136–147, 2013.

[17] K. A. Brogden, “Antimicrobial peptides: Pore formers or metabolic inhibitors in bacteria?,” *Nature Reviews Microbiology*, vol. 3, no. 3, pp. 238–250, 2005.

[18] A. B. Murphy and S. R. Lawrence, “Hemolysis and red blood cell lysis mechanisms,” *Clinical Hemorheology and Microcirculation*, vol. 60, no. 1, pp. 1–13, 2015.

[19] C. A. Schneider et al., “Peptide-induced hemolysis: From molecular interaction to membrane rupture,” *Biophysical Journal*, vol. 102, no. 1, pp. 45–54, 2012.

[20] H. Jiang et al., “Mechanisms of hemolysis induced by cationic peptides: Membrane disruption and pore formation,” *Biochimica et Biophysica Acta (BBA) - Biomembranes*, vol. 1863, no. 4, pp. 183712, 2021.

[21] J. P. Overington et al., “How many drug targets are there?,” *Nature Reviews Drug Discovery*, vol. 5, no. 12, pp. 993–996, 2006.

[22] J. R. Wang and D. Baker, “APIs and data infrastructure for computational biology: Enabling next-generation peptide modelling,” *Bioinformatics*, vol. 39, no. 2, btad019, 2023.

[23] S. Kang et al., “Peptide sequence variation and structural modifications in bioactive peptide databases,” *Briefings in Bioinformatics*, vol. 24, no. 1, bbab437, 2023.

[24] T. Zhao et al., “Machine learning-guided peptide design and hemolytic toxicity prediction,” *Nature Machine Intelligence*, vol. 5, no. 1, pp. 73–85, 2023.

[25] PubMed, “NCBI search strategy for hemolytic peptides (2013–2024)”

[26] G. Wang et al., “APD3: The antimicrobial peptide database as a tool for research and education,” *Nucleic Acids Research*, 2016

[27] The UniProt Consortium, “UniProt: The universal protein knowledgebase in 2023,” *Nucleic Acids Research*, 2023

[28] N. Pirtskhalava et al., “Database of Antimicrobial Activity and Structure of Peptides,” *DAMPD Resource*, SANBI

[29] A. Thomas et al., “CAMP: Collection of antimicrobial peptides,” *Nucleic Acids Research*, 2010

[30] The Apache Software Foundation, “Apache HTTP Server Project,” [Online]. Available: https://httpd.apache.org/. [Accessed: June 2025].

[31] Mozilla Developer Network, “HTML, CSS, and JavaScript for Web Development,” [Online]. Available: https://developer.mozilla.org/. [Accessed: June 2025].

[32] Netcraft, “Usage statistics of PHP and MySQL,” [Online]. Available:

https://news.netcraft.com/. [Accessed: June 2025].

[33] L. Wall, T. Christiansen, and J. Orwant, *Programming Perl*, 4th ed., O'Reilly Media, 2012.

[34] Oracle Corporation, "MySQL: The world's most popular open-source database," [Online]. Available: https://www.mysql.com/. [Accessed: June 2025].

[35] J. Zhang et al., "Comprehensive annotation strategies for peptide-based databases," *Database*, vol. 2022, baac022, 2022.

[36] National Center for Biotechnology Information, "PubMed overview," [Online]. Available: https://pubmed.ncbi.nlm.nih.gov/. [Accessed: June 2025].

[37] S. Sharma et al., "Design and features of peptide databases for functional analysis," *Journal of Proteomics*, vol. 229, pp. 104012, 2020.

[38] D. Piotto et al., "Multiple sequence alignment tools in peptide design: MAP format and beyond," *Briefings in Bioinformatics*, vol. 24, no. 2, bbac061, 2023.

[39] K. van Heel et al., "Cyclic and chemically modified antimicrobial peptides: From structure to function," *Frontiers in Microbiology*, vol. 12, 737010, 2021.

[40] F. Toke, "Antimicrobial peptides: New candidates in the fight against bacterial infections," *Biopolymers*, vol. 80, no. 6, pp. 717–735, 2005.

[41] M. Fjell et al., "Identification of novel synthetic peptide analogs using structure–activity data," *PLOS Computational Biology*, vol. 6, no. 5, e1000823, 2010.

[42] S. Sahoo et al., "Peptide-based therapeutics: Current status and future directions," *Drug Development Research*, vol. 83, no. 4, pp. 368–380, 2022.

[43] M. D. Egbert et al., "Natural vs. synthetic peptides: Distinctions and therapeutic implications," *Expert Opinion on Drug Discovery*, vol. 17, no. 5, pp. 435–446, 2022.

[44] R. R. Mahendran et al., "Insights into peptide activity classification using curated biological databases," *Current Protein & Peptide Science*, vol. 24, no. 1,

pp. 68–77, 2023.

[45] D. E. Vance et al., “The source and derivation of peptide therapeutics: A systematic approach,” *Peptides*, vol. 103, pp. 43–51, 2018.

[46] T. Mehta and R. K. Purohit, “Computational resources for peptide toxicity and hemolysis prediction,” *Briefings in Bioinformatics*, vol. 24, no. 3, bbad051, 2023.

[47] N. Basu and A. Kumar, “Structural annotation and visualization of therapeutic peptides,” *Bioinformatics Advances*, vol. 3, no. 1, vbad025, 2023.

[48] T. Raghava et al., “Peptide search functionalities in therapeutic peptide databases: A usability perspective,” *Journal of Biomedical Informatics*, vol. 125, 104022, 2022.

[49] S. Gautam and T. Raghava, “Advanced search tools in peptide databases for efficient data mining,” *Briefings in Bioinformatics*, vol. 24, no. 1, bbac110, 2023.

[50] G. Landrum, “RDKit: Open-source cheminformatics,” [Online]. Available: https://www.rdkit.org/. [Accessed: June 2025].

[51] S. Sharma et al., “Interactive browsing interfaces for peptide sequence databases,” *Database*, vol. 2022, baac028, 2022.

[52] R. Altschul et al., “Basic local alignment search tool,” *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[53] D. Piotto et al., “Multiple sequence alignment tools in peptide design: MAP format and beyond,” *Briefings in Bioinformatics*, vol. 24, no. 2, bbac061, 2023.

[54] J. Li et al., “Classification of therapeutic peptides based on biological activity using curated databases,” *Molecular Therapy - Nucleic Acids*, vol. 28, pp. 456–467, 2022.

[55] R. Singh and A. Sharma, “Structural classification of peptide drugs: Linear, cyclic, and stereochemical perspectives,” *Expert Opinion on Drug Discovery*, vol. 17, no. 3, pp. 231–243, 2022.

[56] D. Piotto et al., “Peptide modifications and their impact on hemolytic activity: Insights from databases and cheminformatics,” *Biopolymers*, vol. 113, no. 5, e23456, 2023.

[57] G. R. Johnson et al., “Peptide-based toxicity databases: Challenges and opportunities,” *Database*, vol. 2020, baaa016, 2020.

[58] D. S. Wishart et al., “Predicting peptide structure with DSSP and PEPstr,” *Nucleic Acids Research*, vol. 50, no. D1, pp. D650–D655, 2022.

[59] M. Berman et al., “The Protein Data Bank: Enabling structural insights across biological databases,” *Acta Crystallographica Section D*, vol. 76, no. 1, pp. 213–220, 2020.

[60] P. Kapoor and G. Raghava, “An overview of bioassays in peptide databases: Red blood cell lysis and beyond,” *Current Protocols in Bioinformatics*, vol. 71, pp. e97, 2021.

[61] M. Falciani et al., “Peptides in clinical development: Current status and future prospects,” *Current Medicinal Chemistry*, vol. 28, no. 9, pp. 1757–1776, 2021.

[62] B. Zasloff, “Antimicrobial peptides of multicellular organisms,” *Nature*, vol. 415, no. 6870, pp. 389–395, 2002.

[63] T. Zhao et al., “Machine learning-guided peptide design and hemolytic toxicity prediction,” *Nature Machine Intelligence*, vol. 5, no. 1, pp. 73–85, 2023.

[64] P. Pande et al., “Pfeature: A Tool for Computing Wide Range of Protein Features from Sequence and Structure,” *bioRxiv*, 2021. [Online]. Available: https://doi.org/10.1101/2021.01.24.427965.

[65] D. Sharma et al., “HemoPI: A web server for predicting the hemolytic potency of peptides,” *PeerJ Computer Science*, vol. 4, e143, 2018.

[66] T. Chou, “Prediction of protein cellular attributes using pseudo-amino acid composition,” *Proteins: Structure, Function, and Bioinformatics*, vol. 43, no. 3, pp.

246–255, 2001.

[67] P. Refaeilzadeh, L. Tang, and H. Liu, “Cross-validation,” in *Encyclopedia of Database Systems*, Springer, 2009, pp. 532–538.

[68] A. Chou, “Using physicochemical properties to predict protein attributes,” *Journal of Theoretical Biology*, vol. 205, no. 4, pp. 483–491, 2000.

[69] A. Saha and G. Raghava, “Prediction of continuous B-cell epitopes in an antigen using recurrent neural network,” *Proteins: Structure, Function, and Bioinformatics*, vol. 65, no. 1, pp. 40–48, 2006.

[70] C. Shen et al., “Predicting protein-protein interactions based on sequence embedding and conjoint triad features,” *BMC Bioinformatics*, vol. 18, no. 1, 2017.

[71] G. Liu et al., “Predicting anticancer peptides with deep learning and sequence-based descriptors,” *Briefings in Bioinformatics*, vol. 22, no. 5, pp. bbaa162, 2021.

[72] S. Kumar et al., “Peptide property prediction using residue repeat and distribution information,” *Current Bioinformatics*, vol. 15, no. 5, pp. 390–401, 2020.

[73] K. C. Chou, “Prediction of protein cellular attributes using pseudo amino acid composition,” *Proteins: Structure, Function, and Bioinformatics*, vol. 43, pp. 246–255, 2001.

[74] A. M. Reczko and R. Bohr, “Correlation analysis in peptide-based drug discovery,” *Drug Discovery Today: Technologies*, vol. 35, pp. 17–24, 2020.

[75] M. Kandaswamy et al., “Machine learning for peptide-based toxicity prediction using physicochemical properties,” *Journal of Chemical Information and Modelling*, vol. 52, no. 6, pp. 1497–1505, 2012.

[76] E. L. Willighagen et al., “The Chemistry Development Kit (CDK) v2.0: atom typing, depiction, molecular formulas, and substructure searching,” *Journal of Cheminformatics*, vol. 9, no. 1, p. 33, 2017.

[77] T. I. Oprea and J. Gottfries, “Towards predictive ADME models: progress and challenges,” *Current Topics in Medicinal Chemistry*, vol. 1, no. 4, pp. 209–227, 2001.

[78] G. Schneider and U. Fechner, “Advances in the prediction of drug metabolism: quantitative structure–activity relationships and molecular modeling,” *Drug Discovery Today*, vol. 9, no. 6, pp. 265–274, 2004.

[79] M. Kuhn and K. Johnson, *Applied Predictive Modeling*, Springer, 2013.

[80] R. Kohavi, “A study of cross-validation and bootstrap for accuracy estimation and model selection,” in *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, 1995, pp. 1137–1143.

[81] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*, 2nd ed., Springer, 2009.

[82] A. Géron, *Hands-On Machine Learning with Scikit-Learn, Keras, and TensorFlow*, 2nd ed., O’Reilly Media, 2019.

[83] T. Wang et al., “Predicting hemolytic activity of peptides by integrating deep learning with handcrafted features,” *Briefings in Bioinformatics*, vol. 24, no. 1, bbab613, 2023.

[84] J. Demsar et al., “Log-transformed bioactivity data in machine learning: When and why,” *Journal of Cheminformatics*, vol. 10, no. 1, p. 56, 2018.

[85] A. H. Patlewicz et al., “Use of log-transformed data in QSAR modeling: Benefits and drawbacks,” *Regulatory Toxicology and Pharmacology*, vol. 105, pp. 125–134, 2019.

[86] T. Chen and C. Guestrin, “XGBoost: A scalable tree boosting system,” *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016.

[87] S. Raschka, “Model evaluation, model selection, and algorithm selection in machine learning,” *arXiv preprint arXiv:1811.12808*, 2018.

[88] J. Brownlee, *Machine Learning Mastery With Python*, Machine Learning Mastery, 2016.

[89] D. Chicco, L. Warrens, and G. Jurman, “The coefficient of determination R^2 is more informative than SMAPE, MAE, MAPE, MSE and RMSE in regression analysis evaluation,” *PeerJ Computer Science*, vol. 7, e623, 2021.

[90] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An Introduction to Statistical Learning: With Applications in R*, Springer, 2nd ed., 2021.

[91] A. Lundberg and S. Lee, “A unified approach to interpreting model predictions,” *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[92] M. Chen et al., “Ensemble learning for predicting drug-induced toxicity using molecular fingerprints and physicochemical properties,” *Journal of Chemical Information and Modelling*, vol. 61, no. 10, pp. 4982–4993, 2021.