Compilation and Mining
of
Peptide half-life in bodily fluids

Submitted by

Urooj Alam (MT23254)

Under the guidance of

Prof. G.P.S Raghava Head and Professor

In partial fulfilment of the requirements for the degree of Master of Technology in Computational Biology

To

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

May 2025

Certificate

This is to certify that the thesis titled "**Compilation and Mining of Peptide half-life in bodily fluids**" being submitted by Urooj Alam 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.

The results in this thesis have not been submitted in part or full to any other university or institute for awarding any degree/diploma

May 2025

Prof. G.P.S. Raghava

Head & Professor

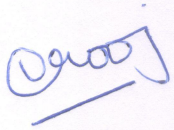
(Department of Computational Biology)

Indraprastha Institute of Information

Technology Delhi- 110020

Acknowledgement

I am deeply grateful to Prof. Gajendra P.S. Raghava at the Indraprastha Institute of Information Technology, Delhi for his invaluable mentorship and for introducing me to this fascinating area of research. His continuous support and insightful guidance have been instrumental in shaping the direction and execution of this project. I would also like to sincerely thank Nishant Kumar (PhD scholar), Pankaj Kumar (PhD scholar), Kunal Choudhary, Ritu Tomer (PhD scholar), Dr. Sumeet whose expertise and encouragement throughout the course of this work have been incredibly helpful. I am also thankful to the Department of Computational Biology and the IT team at IIIT Delhi for providing a conducive research environment and seamless access to the technical resources required for this work. I also wholeheartedly appreciate the encouragement and emotional support from my family and friends, whose belief in me consistently inspired focus and perseverance throughout the course of this thesis.



UROOJ ALAM

MTech. CB (MT23254)

Table of Contents

List of Abbreviation

BBB	Blood-brain barrier
CPP	Cell Penetrating peptides
GIT	Gastrointestinal Tract
GLP-1RA	GLP-1 Receptor Agonist
IM	Intramuscular
IP	Intraperitoneal
IV	Intravenous
MAP	Modification and Annotation in Proteins
NGL	Next Generation Library (3D Viewer)
PEG	Polyethylene Glycol
PDB	Protein Data Bank
SMILES	Simplified Molecular Input Line Entry System
ML	Machine learning
KNN	K-Nearest Neighbor
DPC	Dipeptide Composition
AAC	Amino acid composition
ATC	Atomic Composition

List of Figures

List of Tables

Abstract

The inherent instability of peptides in biological fluids represents a significant challenge for their development as therapeutic agents. To advance the field, we have developed a database of peptide half-life called PEPlife2, which is an updated version of PEPlife. This extensive database comprises 4,486 entries, each comprehensively annotated with experimental conditions, sequence details, modifications, biological activity, and pharmacokinetic parameters. PEPlife2 empowers researchers with powerful data exploration tools, including interactive 3D structural visualization using NGL Viewer, sophisticated sequence-based search functionalities, and a convenient RESTful API. Furthermore, we conducted a thorough analysis of the database to discern key peptide characteristics and subsequently developed machine learning models for half-life prediction. Notably, a K-Nearest Neighbors model developed using dipeptide composition and modification features, yielded excellent predictive performance for modified peptides (validation $R^2 = 0.83$). While an XGBoost model based on amino acid composition was optimal for natural peptides, its accuracy was more modest (validation $R^2 = 0.29$). These results underscore promising avenues for enhancing predictive capabilities through the application of deep learning and large language models (LLMs). PEPlife2 is publicly available at https://webs.iiitd.edu.in/raghava/peplife2/.

Chapter-1

Introduction

1.1 **Definition and Function:**

Therapeutic peptides represent a distinct class of pharmacological agents composed of specific sequences of amino acids whose molecular weight ranges from 500 to 5000 Da [1]. These peptides play vital roles in human physiology and therapeutic applications, functioning as hormones, growth factors, neurotransmitters, ion channel modulators, and antimicrobial agents. Owing to their high specificity and strong binding affinity for cell surface receptors, therapeutic peptides much like biologics such as monoclonal antibodies and therapeutic proteins are capable of initiating intracellular signalling cascades [2–5].

1.2 **Market Trends:**

With approximately 100 peptide-based drugs approved for clinical use and many others in various stages of clinical development, the global market for peptide therapeutics has experienced substantial growth in recent years [6]. A notable milestone in this field is the approval of semaglutide (Rybelsus, Novo Nordisk A/S), the first orally administered glucagon-like peptide-1 receptor agonist (GLP-1RA) for the treatment of type 2 diabetes mellitus (T2DM) and obesity, marking a significant advancement in peptide drug delivery and efficacy [7,8]. As of 2024, semaglutide-based therapies dominate peptide drug sales, with Ozempic (injectable) generating $13.89 billion USD in revenue, followed by Trulicity at $7.13 billion USD, and Rybelsus at $2.72 billion USD. Refer Figure 1 for sales of top top five selling GLP-1 agonists. [9]

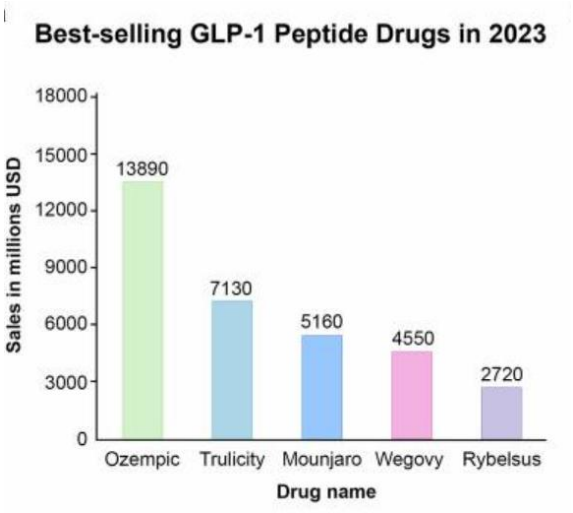


Figure 1: Sales of Top 5 selling GLP-1 agonists

1.3 **Therapeutic Peptides in Disease Treatment:**

1.3.1 **Treatment of Diabetes Mellitus:**

Type 2 diabetes mellitus (T2DM), which predominantly affects middle-aged and older adults, is characterized by impaired glycemic control and acquired insulin deficiency. Peptide-based therapeutics, particularly insulin and glucagon-like peptide-1 receptor agonists (GLP-1RAs), have played a pivotal role in the management of this condition. The endogenous incretin hormone GLP-1, secreted by L-cells in the ileum, exerts its effects through receptors expressed in various tissues, including the gastrointestinal tract, kidneys, lungs, cardiovascular system, central and peripheral nervous systems, and pancreatic β-cells. Upon binding to its receptor, GLP-1 enhances glucose-dependent insulin secretion from β-cells, suppresses glucagon release from α-cells, promotes satiety, and delays gastric emptying [10]. However, native GLP-1 is rapidly inactivated by the enzyme dipeptidyl peptidase-4 (DPP-4), limiting its therapeutic potential. To overcome this limitation, synthetic GLP-1RAs have been developed to resist enzymatic degradation and provide prolonged receptor activation [11].

1.3.2 **Treatment of Cardiovascular Disease:**

Acute myocardial infarction (AMI), a major manifestation of cardiovascular disease (CVD), remains the leading cause of death worldwide, accounting for an estimated 15.9 million fatalities annually. Peptides have emerged as promising therapeutic agents in this domain due to their favorable pharmacokinetic properties, including high solubility, low toxicity, minimal mitogenicity, and ease of structural modification to enhance stability and target affinity [12]. Consequently, various bioactive peptides have been identified that modulate key pathological mechanisms involved in myocardial ischemia-reperfusion injury (MIRI), such as apoptosis, necroptosis, inflammation, and autophagy [13].

1.3.3 **Treatment of Obesity:**

A variety of anti-obesity medications have been developed to induce weight loss, most of which target digestive and absorptive pathways [14]. However, despite their potential efficacy, the clinical utility of these agents is often limited by adverse effects such as dizziness, insomnia, and hypertension, which negatively impact patient adherence [15]. In contrast, peptide-based therapies are gaining increasing attention in obesity treatment due to their favorable pharmacodynamic profiles, lower incidence of side effects, and reduced dosage requirements [16]. Several peptides such as leptin [17], neuropeptide Y (NPY) [18], and adrenomedullin 2 (AM2) [19] have been extensively studied for their roles in appetite regulation, lipid metabolism, and overall energy homeostasis.

1.3.4 **Treatment of Cancer:**

Anticancer peptides (ACPs) are short sequences of amino acids that exhibit selective cytotoxicity toward cancer cells, often through distinct amino acid motifs and structural features [20]. Compared to conventional therapies such as monoclonal antibodies and small-molecule drugs, ACPs offer several advantages, including high tumor selectivity, enhanced tissue penetration, and ease of chemical modification [21]. Numerous peptide-based vaccines and therapeutic candidates are currently undergoing clinical evaluation. One notable example is CIGB-300, an amidated, disulfide-cyclized undecapeptide linked to the TAT cell-penetrating peptide via a β-alanine spacer. CIGB-300 exerts its anticancer effects by inhibiting casein kinase 2 (CK2)-mediated phosphorylation, ultimately inducing apoptosis in cancer cells. Promising results from clinical studies have demonstrated its efficacy in patients with non-small cell lung cancer and cervical cancer [22–24].

1.3.5 **Treatment of Infections:**

Antimicrobial peptides (AMPs), typically consisting of 10–50 amino acids, are characterized by a net positive charge ranging from +2 to +11 and a high proportion of hydrophobic residues, often comprising approximately 50% of their structure [25–27]. Several AMPs such as pexiganan, omiganan, and the synthetic peptide LTX-109—are currently undergoing clinical evaluation for the treatment of various bacterial infections. Pexiganan, a 22-amino-acid derivative of the amphibian peptide magainin, functions by disrupting bacterial membranes and has been assessed in Phase III clinical trials as a topical agent for bacterial infections associated with diabetic foot ulcers [28,29].

1.3.6 **Treatment of Neurological diseases:**

Traditional small-molecule drugs often fail to effectively address the underlying causes of neurological disorders due to the intricate nature of neural processes, the typically irreversible progression of neurological degeneration [30,31], and the restrictive properties of the blood-brain barrier (BBB) [32]. These limitations frequently lead to suboptimal outcomes and unintended neurological side effects. Most pharmacological treatments for neurological conditions primarily function by modulating neurotransmitter activity. For instance, galantamine marketed as Reminyl, a reversible cholinesterase inhibitor commonly prescribed for the management of Alzheimer's disease (AD) [33]. Similarly, fluoxetine, a selective serotonin reuptake inhibitor (SSRI) sold under the brand name Prozac, is widely used to treat depression, particularly in adolescents [34].

1.4 **Challenges in peptide therapeutics:**

Peptide-based therapeutics face several developmental challenges, including poor absorption, rapid enzymatic degradation, low systemic stability, and swift excretion. These factors significantly constrain their therapeutic potential and limit overall efficacy.

1.4.1 **Enzymatic degradation and Renal Clearance:**

Due to the presence of amide bonds in their structure, peptides are highly susceptible to degradation by proteases and peptidases. This degradation is facilitated by both luminally secreted enzymes—such as pepsin, elastase, trypsin, and chymotrypsin and brush-border membrane-bound enzymes, including endopeptidases, aminopeptidases, and carboxypeptidases. The human body contains over 550 putative proteases, underscoring the challenge of maintaining peptide stability in vivo [35].

In addition to enzymatic degradation, peptides often exhibit short plasma half-lives, primarily due to rapid metabolic breakdown and efficient renal clearance. Hydrophilic peptides with molecular weights below 25 kDa can easily pass through the glomerular filtration barrier and are rapidly eliminated by the kidneys. Unlike many small molecules, peptides are not readily reabsorbed in the renal tubules, resulting in high renal clearance. Consequently, in patients with impaired kidney function, dosage adjustments are necessary to prevent drug accumulation, which may lead to toxicity or diminished therapeutic efficacy [36].

1.4.2 **Low oral bioavailability:**

Oral administration remains the most common and preferred route for drug delivery due to its convenience and high patient compliance. Most drug absorption occurs in the gastrointestinal tract (GIT) [37]. However, macromolecular drugs such as peptides often exhibit poor oral bioavailability due to several limiting factors, which can be broadly classified into two categories. The first category pertains to the physicochemical properties of the drug itself, including molecular size, charge, polar surface area, lipophilicity, and the ability to donate or accept hydrogen bonds. The second category involves host-related biological barriers, such as hepatic metabolism, intestinal transport mechanisms, gastrointestinal transit time, efflux pumps, and the activity of luminal and mucosal enzymes [38].

1.4.3 **Rapid systemic clearance:**

The short plasma half-lives of peptides and proteins are largely due to their rapid renal clearance attributable to their hydrophilic nature and relatively small size as well as enzymatic degradation by proteases found in the blood, liver, and kidneys. To address this, it is essential to identify the specific enzymes responsible for degrading a given peptide during systemic circulation. In addition to mapping the tissue distribution of peptidases and proteases, understanding their substrate specificity is particularly important. Developing strategies to extend plasma half-life could improve the pharmacokinetic properties of current therapeutics and potentially broaden their clinical applications [39].

1.4.4 **Stability and Delivery challenge:**

Despite their high biological selectivity and favourable safety profiles, peptides often face challenges related to short plasma half-lives and poor proteolytic stability. These limitations can lead to the gradual loss of secondary structure, ultimately diminishing their biological activity over time [40]. Cell-penetrating peptides (CPPs), also known as cell transduction domains (CTDs), represent a novel class of peptides with the unique ability to cross cellular membranes. Owing to their efficient intracellular delivery capabilities, CPPs are promising vectors for transporting a wide range of therapeutic cargo, including peptides, DNA, siRNA, and small-molecule drugs [41–44].

CPPs can bind to target molecules through covalent or noncovalent interactions, such as hydrophobic or electrostatic forces, facilitating their transport across biological barriers [45]. Additionally, peptide stability against enzymatic degradation can be enhanced by chemically reinforcing structural motifs like α-helices. One such approach is hydrocarbon stapling, which helps maintain helical conformation and improves resistance to proteolysis [46]. These strategies offer promising solutions to the persistent challenges of peptide stability and systemic delivery.

1.5 **Strategies to enhance the stability of peptides:**

1.5.1 **Cyclization:**

Cyclization is a widely used and relatively straightforward chemical strategy that enhances several key properties of linear peptides and peptidomimetics, including increased cell permeability, improved metabolic stability, and greater target selectivity [47]. The conformational rigidity of cyclic peptides reduces the entropic cost of target binding, thereby facilitating more efficient and selective interactions. In contrast, the structural flexibility of linear peptides increases the risk of off-target effects due to non-specific binding [48]. Various cyclization methods such as end-to-end (head-to-tail), side-chain-to-side-chain, and end-to-side-chain cyclization offer distinct advantages for optimizing the pharmacological and biophysical properties of peptides.

1.5.2 **D-amino acids:**

Endogenous proteins and peptides are composed primarily of L-amino acids, which are readily recognized and degraded by proteolytic enzymes. In contrast, D-amino acids exhibit greater resistance to enzymatic degradation due to their stereochemical configuration, which is not easily recognized by these enzymes. Therefore, substituting L-amino acids with D-amino acids in therapeutic peptides can significantly enhance their proteolytic stability [49].

1.5.3 **Peptoids:**

Peptidoids, a class of peptidomimetics composed of N-substituted glycines, exhibit enhanced proteolytic stability due to the absence of amide bonds in their backbone. Their structural similarity to peptides also enables efficient synthesis through solid-phase methodologies [50].

1.5.4 **N-methylation:**

Alkylation of the nitrogen atom in peptide amide bonds is a key chemical modification that enhances the pharmacokinetic properties such as absorption, half-life, and bioavailability and overall drug-like characteristics of therapeutic peptides. Among various N-alkylation strategies, N-methylation is the most widely employed due to its versatility and ease of synthesis [51].

1.5.5 **Delivery system for peptides:**

There is increasing interest in utilizing advanced delivery systems to protect therapeutic peptides from enzymatic degradation, rather than relying solely on chemical modifications. Various drug carriers such as nanoparticles, liposomes, dendrimers, micelles [52], polyelectrolyte complexes [53], hydrogels [54], and biomimetic or bio-derived nanocomposites [55] have been explored for this purpose. Among these, nanomaterials have shown particular promise for the targeted delivery of peptides and proteins. The effectiveness of nanoparticle-based delivery systems is closely linked to their physicochemical properties, including particle size, shape, surface charge, and porosity, all of which influence the structure–function relationship essential for efficient drug transport and release [56].

1.5.6 **Protecting N- and C-terminus:**

Peptide degradation often occurs at the N- and C-termini through the action of various proteolytic enzymes present in the blood/plasma, liver, and kidneys, including exopeptidases, aminopeptidases, and carboxypeptidases. Modifying the terminal ends of peptides can significantly enhance their stability. In particular, N-acetylation and C-amidation have been shown to improve resistance to enzymatic degradation in multiple cases [57][58]. For example, N-terminal acetylated analogs of somatostatin exhibit substantially greater stability compared to the native peptide [59].

1.5.7 **Modification of Amino acids:**

Modifications involving natural amino acids can enhance peptide stability by introducing steric hindrance or disrupting enzyme recognition sites [60].

1.5.8 **Conjugation to macromolecules:**

Modifications using natural amino acids can improve peptide stability by introducing steric hindrance or interfering with enzyme recognition sites [60]. In addition, traditional polymer

conjugation strategies have been developed to enhance the half-life and reduce the immunogenicity of proteins and peptides with poor systemic pharmacokinetics (PK). Polymers provide steric shielding that protects therapeutic molecules from opsonization, proteolytic degradation, and antidrug antibody interactions. Furthermore, the increased molecular size resulting from polymer attachment reduces renal clearance. Together, these effects significantly prolong the circulation half-life of polymer–protein and polymer–peptide conjugates, allowing serum concentrations to remain within the therapeutic window for extended durations [61].

Therefore, the application of such chemical and physical modifications presents an effective strategy to improve the stability and pharmacokinetic profiles of peptide-based therapeutics.

1.6 **PEPlife2 and Other Databases**:

Peptide half-life is defined as the time required for 50% of a peptide's concentration to degrade or be eliminated from a biological system. This critical pharmacokinetic parameter plays a pivotal role in determining dosing frequency, therapeutic efficacy, and the overall design of peptide-based drugs. Several factors influence peptide half-life, including the biological matrix, the presence of proteolytic enzymes, chemical and physical modifications, peptide concentration, chirality, and the type of assay employed (in vitro or in vivo).

Specialized databases such as PEPlife [62] and PepTherDia [63] have been established to facilitate peptide research by offering curated data on peptide half-lives and associated pharmacological attributes. Among them, PEPlife stands out for its comprehensive collection of experimentally validated half-life data. Launched in 2016, PEPlife includes 2,229 entries encompassing 1,193 distinct peptides and 37 unique proteins, with data compiled from studies conducted under diverse biological conditions. Additionally, the database has served as a foundation for the development of computational models aimed at predicting peptide stability and degradation kinetics [64].

In contrast, PepTherDia centers on more than 100 approved peptide therapeutics, offering detailed information on terminal half-life, administration routes, structural features, and clinical applications. Although it serves as a valuable resource for understanding the pharmacokinetic characteristics of marketed peptides, the half-life data it provides are not always derived from direct experimental measurements, but are frequently sourced from clinical trials or regulatory documentation.

Although PEPlife has been a valuable tool, its lack of updates since its initial release in 2016 created a need for more current and comprehensive resources. This gap has been filled by the development of PEPlife2, a significantly enhanced version of the original database. PEPlife2 features 4,486

entries and provides extensive information on peptide sequences, chemical composition, proteases involved, assay types, test samples, and notably, route of administration and physical modifications, details that were absent from the initial version. With its broader coverage and improved functionality, PEPlife2 represents a timely and robust resource for advancing research in peptide and protein therapeutics.

Chapter-2

Compilation of peptides half-life in bodily fluids

MATERIALS AND METHODS

2.1 Data Curation:

An updated version of the PEPlife database, first established in 2015, now includes data published from 2016 onward, expanding the original content with newly available information. Full-text records of granted patents were obtained from The Lens, and relevant literature was sourced from PubMed using the query "((half-life[Title/Abstract]) AND (peptide[Title/Abstract]) NOT Review," targeting studies from May 2016 to November 2024. Only peptides with experimentally determined half-life values were considered for inclusion. The curated dataset comprises both natural and modified peptides, along with associated experimental parameters such as test sample, origin, biological media, physical modifications, chemical modification, sequence and reported half-lives. As of June 2024, 1243 articles were initially retrieved. Following the exclusion of irrelevant and review articles, approximately 950 publications were screened in detail, resulting in the selection of 449 articles for final curation. Additionally, 18 relevant patents identified through The Lens database were manually reviewed and incorporated into the dataset.

2.2 Database Architecture and Web Interface:

The PEPlife2 database was developed using the LAMP (Linux-Apache-MySQL-PHP) technology stack. The backend infrastructure was established with Apache (version 2.4.7) serving as the HTTP server and MySQL (version 5.7.31) as the database management system. The frontend interface, compatible with desktops, tablets, and smartphones, was implemented using HTML5, CSS3, PHP (version 7.3.21), and JavaScript (version 1.8). A user-friendly web interface was constructed using PHP to facilitate interaction with the database. The system architecture of PEPlife2 is illustrated in Figure 2.

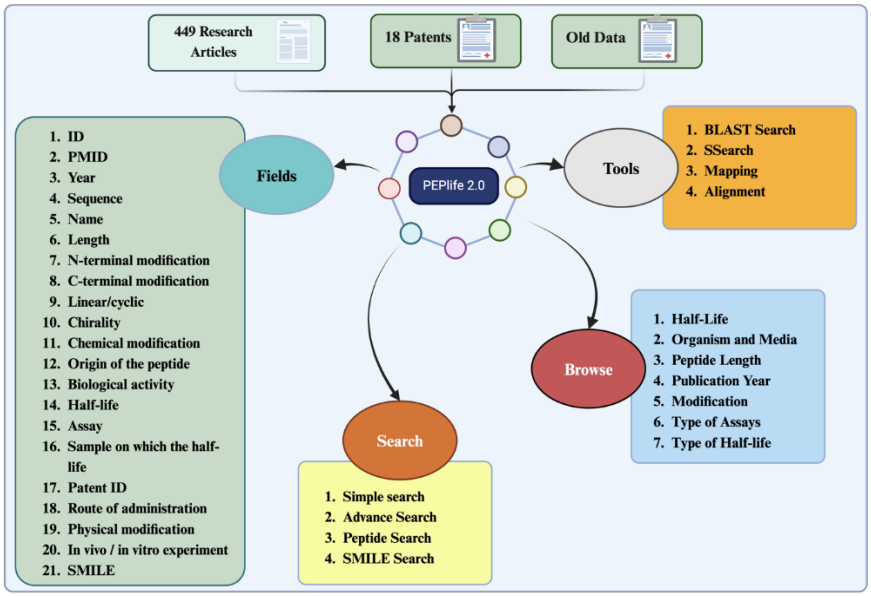


Figure 2: Architecture of PEPlife2

2.3 Data Content of PEPlife2 Database:

PEPlife2 is a well-curated resource that offers detailed information on the half-life of peptides and proteins, along with related biochemical properties and experimental parameters. The database includes fields such as: (1) PEPlife2 ID: Unique ID assigned to each entry, (2) PMID: PubMed ID, (3) Year: Year of Publication, (4) Sequence: Peptide or Protein Sequence, (5) Name of peptide, (6) Chirality: L/D/Mix, (7) N-terminal modification: Modification at N terminal of peptide, (8) C-terminal modification: Modification at C terminal of peptide, (9) Linear or Cyclic structure, (10) Length of peptide, (11) Chemical modification: Modification other than terminal modification, (12) Origin of the peptide, (13) Biological activity of the peptide, (14) half-life, (15) Units of half-life, (16) sample on which the half-life was tested, (17) Assay used to measure half-life, (18) Patent ID if applicable, (19) Route of administration of peptide (IV/IM/IP/Oral), (20) Physical modification (Hydrogel/Nanosystem), (21) In-vitro/In-vitro (Experiment conducted in-vitro/in-vitro), (22) SMILES: SMILES format for peptide, (23) Incubation time of peptide, (24) Concentration of peptide used, (25) Protease present.

The peptide and protein sequences in PEPlife2 were further utilized to generate secondary data such as tertiary structures and SMILES representations. Structural information in the database is depicted using 2D images. Where available, experimentally determined structures were obtained by matching sequences with entries in the Protein Data Bank (PDB). If no PDB match was found, peptide structures longer than five amino acids were predicted using the PEPstrMOD webserver [65]. Peptides with complex modifications (e.g., heparosan, Streptavidin) were excluded from

structure prediction due to the unavailability of corresponding force field parameters. For sequences exceeding 40 residues, I-TASSER [66] was employed for structure prediction. In the case of peptides shorter than five amino acids, a linear structure was assumed, using default dihedral angles (ϕ and $\psi = 180°$), followed by energy minimization and molecular dynamics (MD) simulations. Secondary structural information for all peptides was derived from their predicted or retrieved tertiary structures using DSSP (Define Secondary Structure of Proteins) software [67], which categorizes residues into one of eight classes: Beta-bridge (B), Coil (C), Extended strand (E), 3/10 helix (G), Alpha-helix (H), Pi-helix (I), Bend (S), and Turn (T).

2.4 Implementation of Tools:

2.4.1 Search tool:

PEPlife2 offers a user-friendly search interface comprising several modules for retrieving data. The **Basic Search** module enables users to perform straightforward queries using specific keywords across any individual field. Users can search one keyword at a time based on parameters such as “PMID”, “year”, “peptide name”, “sequence”, “terminal modifications”, “biological activity”, and “nature of the peptide”.

For more complex queries, the **Advanced Search** module is available. It allows users to combine multiple fields simultaneously using Boolean operators such as “AND”, “OR”, and “NOT” for refined and comprehensive searches.

The **Peptide Sequence Search** feature supports both “identical” and “subsequence” searches. Identical search retrieves exact sequence matches, whereas subsequence search returns all peptides that contain the specified sequence fragment.

Additionally, the **SMILES Search** module enables atom-level structural searches using the SMILES notation of a peptide. Users can search by “substructure”, “exact structure”, “exact fragment”, or “superstructure”, offering high precision in structure-based retrieval.

2.4.2 Browsing tool:

PEPlife2 also features an intuitive browsing tool that enables users to explore the dataset based on several predefined categories. Users can browse peptides according to their half-life, with entries grouped into specific time ranges as represented in the "hl" (half-life) column. Another browsing option is based on the “organism or media” in which the peptide was tested, including sources such as cat, mare, human, rat, monkey, and others. The tool also allows navigation by “year of publication”, enabling users to access peptides introduced or studied in specific years. Peptides can also be filtered by their “length”, facilitating comparisons across sequences of varying sizes. Lastly,

browsing by “modifications” is supported, covering “N-terminal”, “C-terminal”, and “chemical modifications”, helping users quickly identify structurally modified peptides.

2.4.3 Analysis tool:

To support similarity-based exploration, PEPlife2 incorporates a range of bioinformatics tools designed for both sequence and structural comparisons. Among these, the **Basic Local Alignment Search Tool (BLAST)** and the **Smith-Waterman algorithm** [68][69] are utilized to detect sequence similarities between user queries and database entries. Users can input peptide sequences in **FASTA format**, choosing either default or user-defined parameters, after which the system automatically executes a BLAST search against the database. For more refined local alignments, a dedicated **Smith-Waterman-based engine** enhances the precision of similarity analysis. The **Peptide Mapping** tool further allows identification of peptide segments within larger protein sequences, facilitating accurate localization of known peptides. A **Sequence Alignment** function enables direct comparison between submitted sequences and existing PEPlife2 entries. Additionally, the **Structural Alignment** module allows users to upload **PDB files** and perform three-dimensional structure comparisons against peptide structures stored in the database, offering deeper insights into structural relationships.

2.5 Download:

PEPlife 2.0 offers users convenient access to its full dataset through various downloadable formats like CSV or Excel. Peptide sequences, including both Natural and modified versions, can be retrieved in multi-FASTA format and structural data are available as PDB files which includes both experimentally validated and predicted structures.

2.6 API:

To streamline data access and support automated workflows, PEPlife 2.0 now includes a **RESTful API** for programmatic retrieval. This feature enables seamless integration with external tools by allowing users to query the database using straightforward URL requests. Data can be fetched based on criteria such as **Linear/Cyclic**, **Organism/media**, or **peptide sequence category**, and is delivered in JSON format. This approach eliminates the need for manual downloads and gives users full control over how the data is parsed and utilized in their own systems.

Figure 3 illustrates the PEPlife2 webserver homepage, which features several key modules including the “Search,” “Browse,” and “Analysis” modules, along with additional tools like “Download” and “API” options for seamless access to the complete dataset.

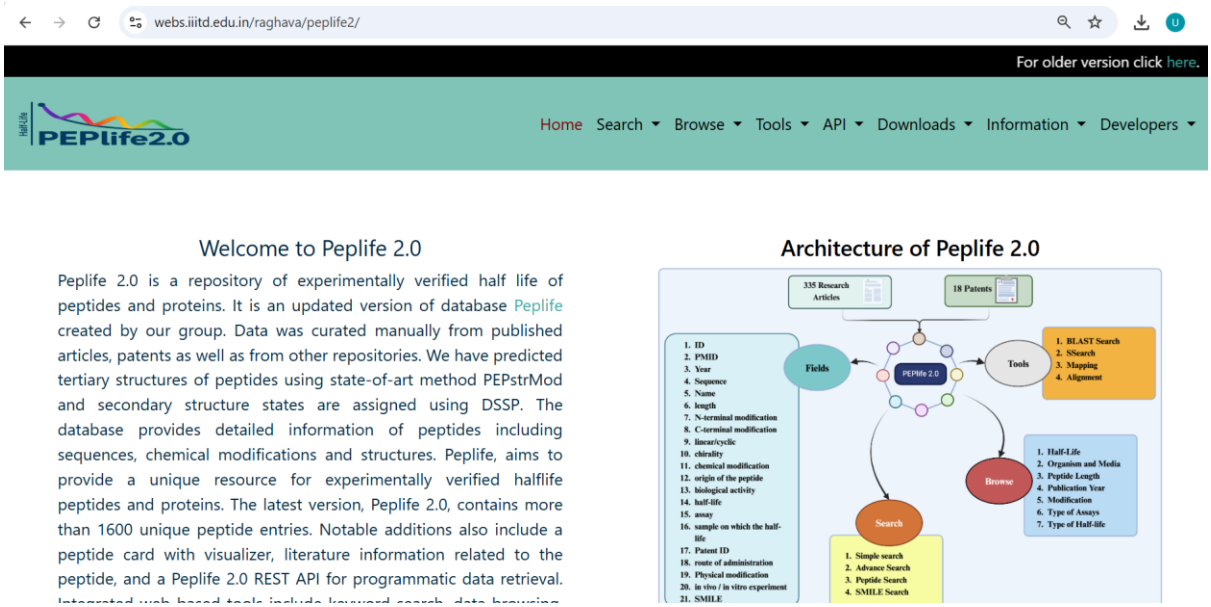


Figure 3: PEPlife2 Webserver Homepage

RESULTS AND DISCUSSION

The PEPlife2 database (https://webs.iiitd.edu.in/raghava/peplife2/) is an enhanced version of the original PEPlife database. It includes a total of **4,486** entries - 2,229 entries from the previous PEPlife, along with 2,183 new peptide entries and 74 miscellaneous entries. This total encompasses 2,046 peptides sourced from 449 research publications and 211 peptides derived from 18 patents. The database features peptides of varying lengths, from single-residue peptides to those exceeding 54 residues. The largest group consists of peptides between 1 and 10 residues, followed by those in the 11 to 21 residue range, and then peptides with 33 to 43 residues. The smallest group is comprised of peptides ranging from 44 to 54 residues. Most of the peptides in PEPlife2 are “linear”, with 3,728 entries, while “cyclic” peptides account for 655 entries. Additionally, terminal modifications are frequently observed, with 2,104 peptides exhibiting “N-terminal modifications” and 1,481 featuring “C-terminal modifications”. Various techniques, including “Mass spectrometry”, “HPLC”, and “ELISA”, were used to measure the half-life of the peptides. A comprehensive breakdown of peptides and proteins by their length, Chirality of peptide (or mixed configuration), structural type, organism or media and the assays employed to determine their half-life can be found in Figure 4 and 5.

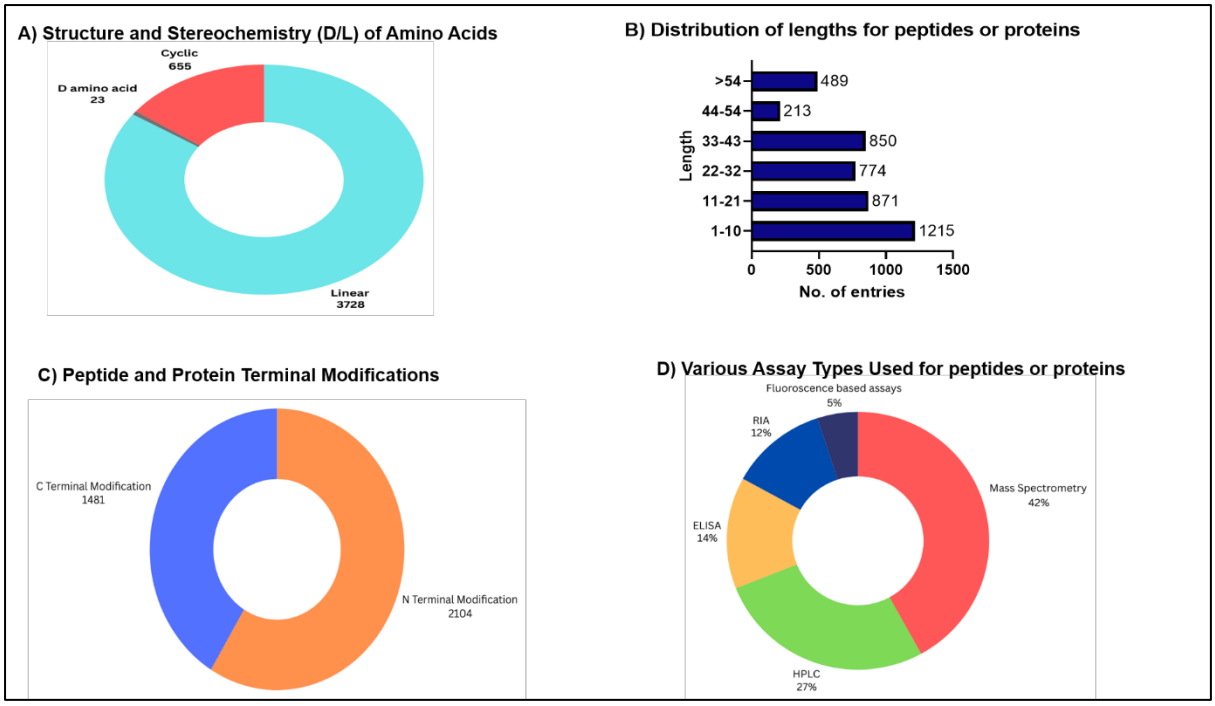


Figure 4: Distribution of peptides and peptides based on (A) Chirality of amino acid and Structure (B) Length of peptides and proteins (C) N- and C-terminal modifications (D) Types of Assay used

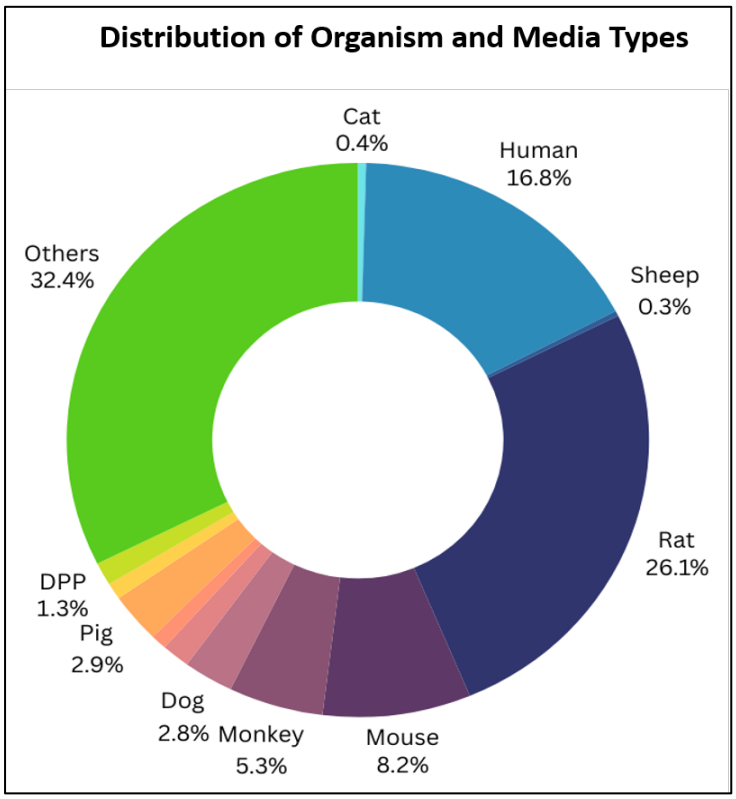


Figure 5: Distribution of organism or media used

Additionally, a comprehensive breakdown of the distribution of unique peptides and their corresponding protein lengths is provided in Figure 6 below.

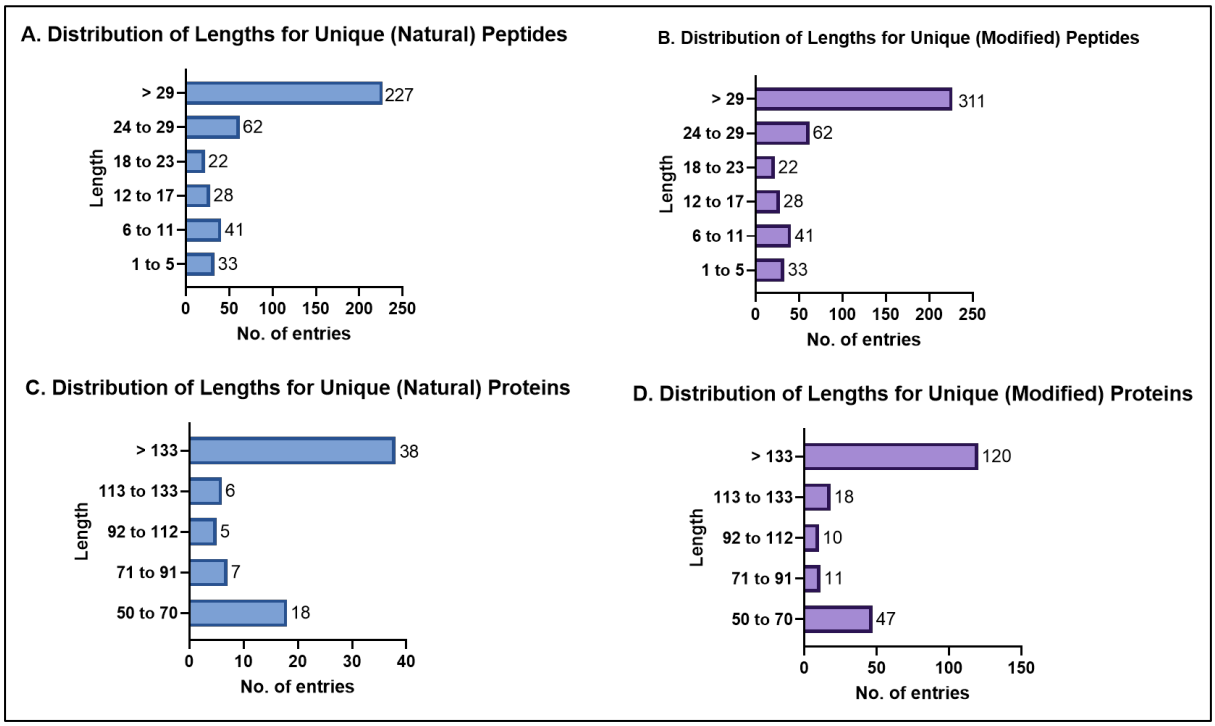


Figure 6: Length Distribution for Unique peptide and Unique proteins

The database compiles a comprehensive array of chemical modifications aimed at enhancing the stability and extending the half-life of peptides and proteins. Figure 7 illustrates the influence of these modifications on peptide half-life. Among the included modifications are N-methylation, PEGylation, amidation, acetylation, incorporation of unnatural amino acids, D-amino acid substitutions, fatty acid attachment (commonly at lysine residues), Fc conjugation, cyclization, and more.

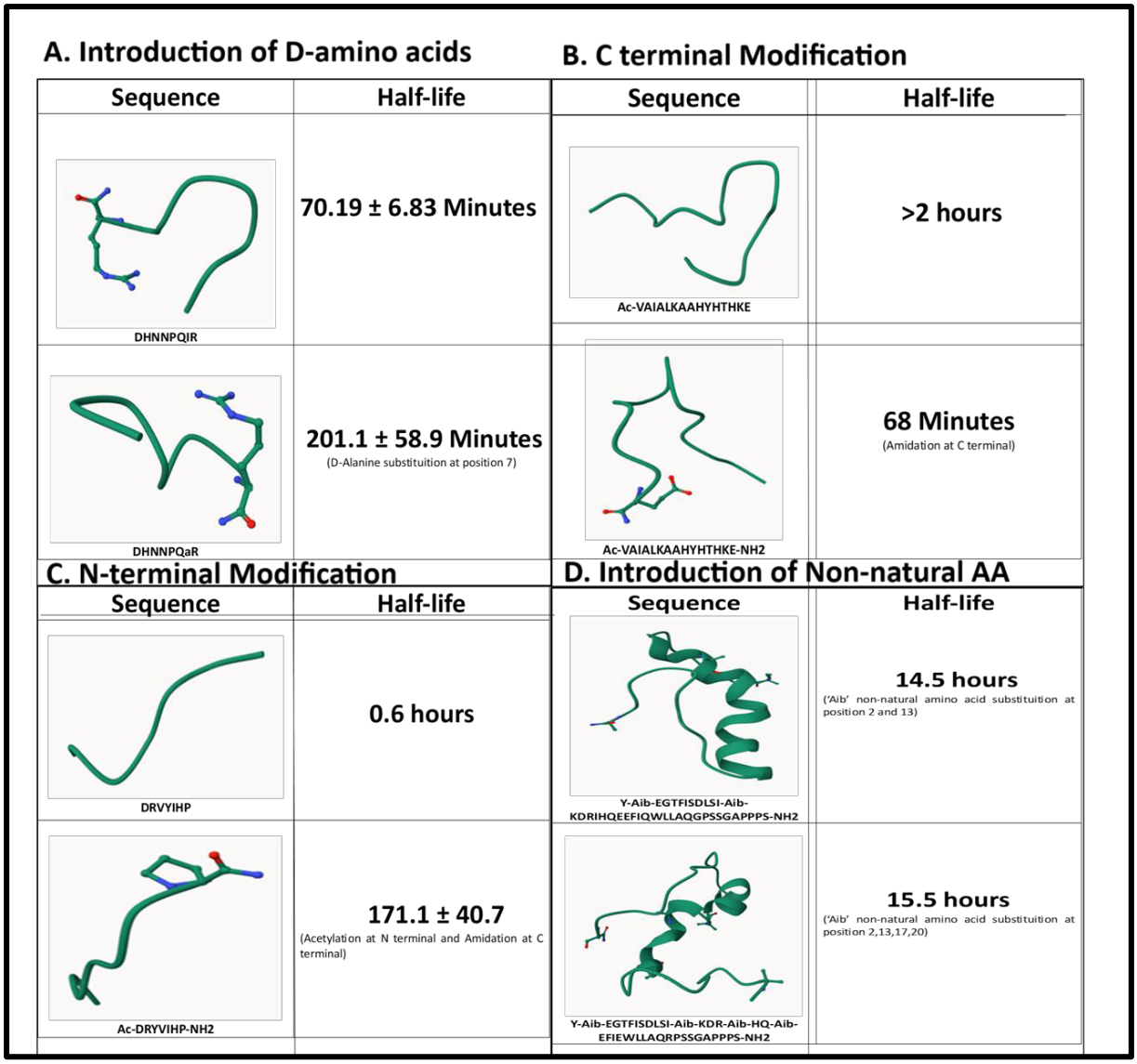


Figure 7: Effect of Chemical modifications on Peptide Half-life

For instance, the peptide DR8 (sequence: DHNNPQIR-NH_2) exhibits a half-life of 70.19 ± 6.83 minutes. When D-Alanine is substituted at position 7, its half-life increases significantly to 201.08 ± 58.86 minutes. Similarly, the peptide Ang(1–7) with sequence DRVYIHP shows enhanced stability when modified at both termini - N-terminal acetylation and C-terminal amidation—yielding the modified peptide Ac-DRVYIHP-NH_2 (Ang-AA), which displays an extended half-life.

Another example involves the incorporation of the unnatural amino acid Aib (α-aminoisobutyric acid) into the peptide SEQ ID NO 35. SEQ ID NO 35 with sequence Y-Aib-EGTFISDLSI-Aib-KDRIHQEEFIQWLLAQGPSSGAPPPS has a half-life of 14.5 hours. With additional Aib substitutions (resulting in Y-Aib-EGTFISDLSI-Aib-KDR-Aib-HQ-Aib-EFIEWLLAQRPSSGAPPPS), the half-life increases to 15.5 hours.

In a case of physical modification, the AAP1 peptide conjugated to a gold nanorod (GNRs-AAP1-1-Cy5) demonstrates an enhanced half-life of 0.18 hours due to the presence of the nanomaterial.

To enable residue-level annotation of such modifications, the database employs the MAP (Modification and Annotation in Proteins) format [70]. This system allows for precise tagging of individual residues to denote chemical modifications, unnatural amino acids, mutations, cyclization, conjugation to macromolecules, and post-translational changes. For example if peptide has terminal modifications Acetylation and Amidation both then MAP format will be – LPFFGKH{nt:Acet}{ct:Amid}, if there is Fatty acid (C16) attachment then it can be written as LPFFGK{conj:palm}H{nt:Acet}{ct:Amid} and if there is unnatural amino acid substituitions then sequence can be written as FG{nnr:Aib}HKKAG.

COMPARISON WITH PREVIOUS PEPlife VERSION:

The original PEPlife database, released in 2016 and covering data up to 2015, contained 2,229 entries corresponding to 1,193 unique peptides. Each record included comprehensive details such as the peptide's name, sequence, length, type of chemical modification, half-life, the assay used for half-life determination, and associated biological activity and classification. However, this earlier version lacked the capacity to accommodate updates related to peptide half-life, assay variations, sample types, physical modifications, and newer chemical modifications. As a result, an updated version became necessary. The new release PEPlife2 addresses these limitations by incorporating data published between 2016 and 2024. This expanded dataset is derived from 449 research articles and 18 patents, significantly increasing the scope and depth of the database. PEPlife2 introduces several key enhancements not present in the original version. These include: physical modifications (now a total of 19) (e.g., Gold Nanorods, Hydrogels), different test samples (now a total of 30), different assay used (now a total of 130). It also contain different types of half-life mentioned alongwith half-life if applicable. Here below is the detailed PEPlife2 statistics including unique peptides and proteins information for different fields in Figure 8 (https://www.biorender.com/).

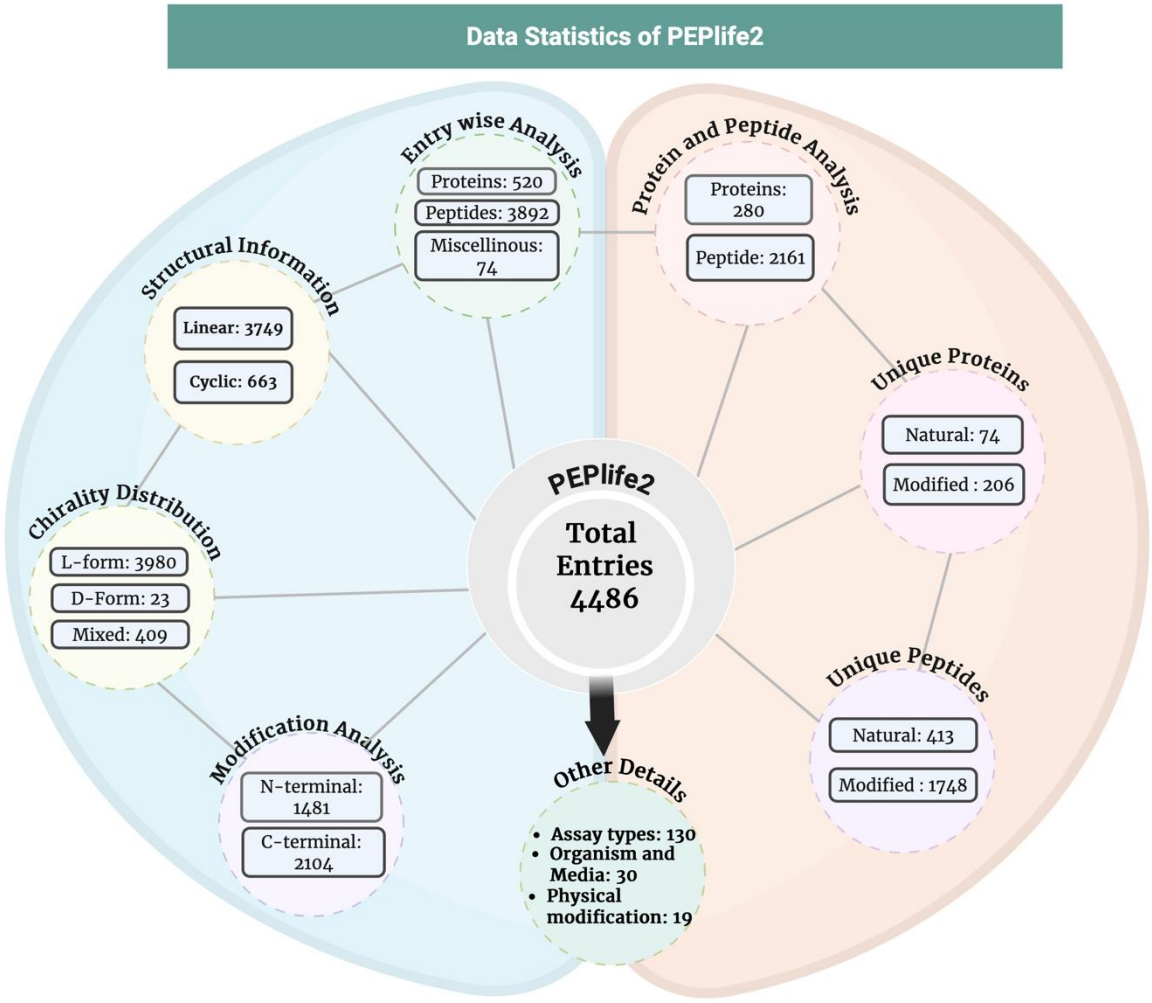


Figure 8: The Complete Stats of PEPlife 2.0

PEPlife 2.0 introduces several major functional upgrades that significantly improve user experience and accessibility. One of the standout features is the integration of the NGL viewer, which enables interactive 3D visualization of peptide structures directly within the database interface. Each peptide entry is presented in a detailed peptide card, which includes essential information such as the PubMed ID (PMID), peptide/protein name, sequence, length, biological nature, and origin. Additional structural data are also provided, including the SMILES representation and a downloadable PDB file for further analysis.

To support automated data workflows, PEPlife 2.0 now offers a RESTful API, allowing users to programmatically retrieve data through simple URL-based queries. Users can filter entries based on parameters such as peptide type (linear or cyclic), organism or media, or sequence category, with results returned in JSON format. This streamlined access eliminates the need for manual downloads and enables seamless integration with external tools and pipelines.

Together, these enhancements make PEPlife 2.0 a far more comprehensive and user-friendly resource for researchers studying peptide stability and chemical modifications, significantly advancing beyond the capabilities of the original version.

COMPARISON WITH OTHER DATABASES

Currently there are only 2 dedicated database for peptide half-life which include PEPlife and PEPTherDia. PEPlife2 is the only database that provide comprehensive information about experimentally determined half-life along with other details like sequence, length, terminal modifications, chemical modifications, origin, test sample, protease, assay, incubation time, concentration, physical modification, etc. Below is the Table 1 for comparison with different databases.

Table 1: Comparison of PEPlife2 with other database

	PEPlife	PEPTherDia	PEPlife 2.O
Primary focus	Experimentally derived peptide half-life	Approved peptide therapeutics and diagnostics	Experimentally derived peptide half-life
Number of entries	2,229 entries	105 approved peptide drugs and diagnostics	4,486 entries
Half-life data	Experimentally determined half-lives, assay-specific	Terminal half-lives compiled from literature and DrugBank	Experimentally determined half-lives, assay-specific
Peptide Sequence	Natural + Modified	Natural + Modified	Natural + Modified
Modifications included	Extensive	Limited (Only modifications present in approved drugs)	More extensive
Structure data	PDB files, SMILES, 2D visualization with pre-requisite Jalview tool	if available on DrugBank	PDB files, SMILES, 3D visualization with no pre-requisite NGL viewer
API access	No	No	Yes
Download options	Manual download of drug/peptide records	Manual download of drug/peptide records	Downloadable PDB, CSV, Excel, JSON formats and MAP,natural,modified sequence file
Sequence alignment	Yes	No	Yes
Data sources	PubMed (335), DrugBank, patent lens (16)	DrugBank, literature, company websites, clinical sources	PubMed (449), patent lens (18)

Chapter-3

Mining of Peptides and Proteins

Natural peptides, filtered to include lengths between 5 and 50 residues and sourced from mammalian blood with half-lives ranging from 20 seconds to 24 hours, were analyzed for their amino acid composition. The correlation between each amino acid and peptide half-life was calculated using Pearsonr revealing that **Alanine** had the strongest positive correlation, while **Valine** showed the weakest, as summarized in Table 2.

Table 2: Correlation of Amino acid composition (AAC) with half-life for Natural peptide

Amino acid	Correlation with Half-life
A	0.20
C	0.039
D	0.096
E	0.090
F	-0.13
G	-0.20
H	0.10
I	0.079
K	-0.094
L	-0.047
M	0.0078
N	0.11
P	-0.11
Q	-0.031
R	0.039
S	-0.052
T	0.072
V	-0.0051
W	0.12
Y	-0.014

Natural peptides, filtered to include lengths between 5 and 50 residues and sourced from mammalian blood with half-lives ranging from 20 seconds to 24 hours, were analyzed for their Amino acid frequency. The correlation between each amino acid and peptide half-life was calculated using Pearsonr revealing that **Tryptophan** had the strongest positive correlation, while **Methionine** showed the weakest, as summarized in Table 2.

Table 3: Correlation of Amino acid frequency (AAF) with half-life for Natural peptide

Amino acid / Length	Correlation with Half-life
A	0.31
C	0.026
D	0.17
E	0.16
F	-0.15
G	-0.18
H	0.077
I	0.21
K	-0.048
L	0.044
M	0.0097
N	0.17
P	-0.19
Q	0.04
R	0.13
S	-0.039
T	0.20
V	-0.097
W	0.38
Y	0.12
Length	0.12

The average amino acid composition (%) of peptides with long and short half-lives was compared. It was observed that peptides with longer half-lives were enriched in **alanine, glycine, and serine**, while those with shorter half-lives showed higher levels of **arginine, serine, and glycine**, as illustrated in Figure 9.

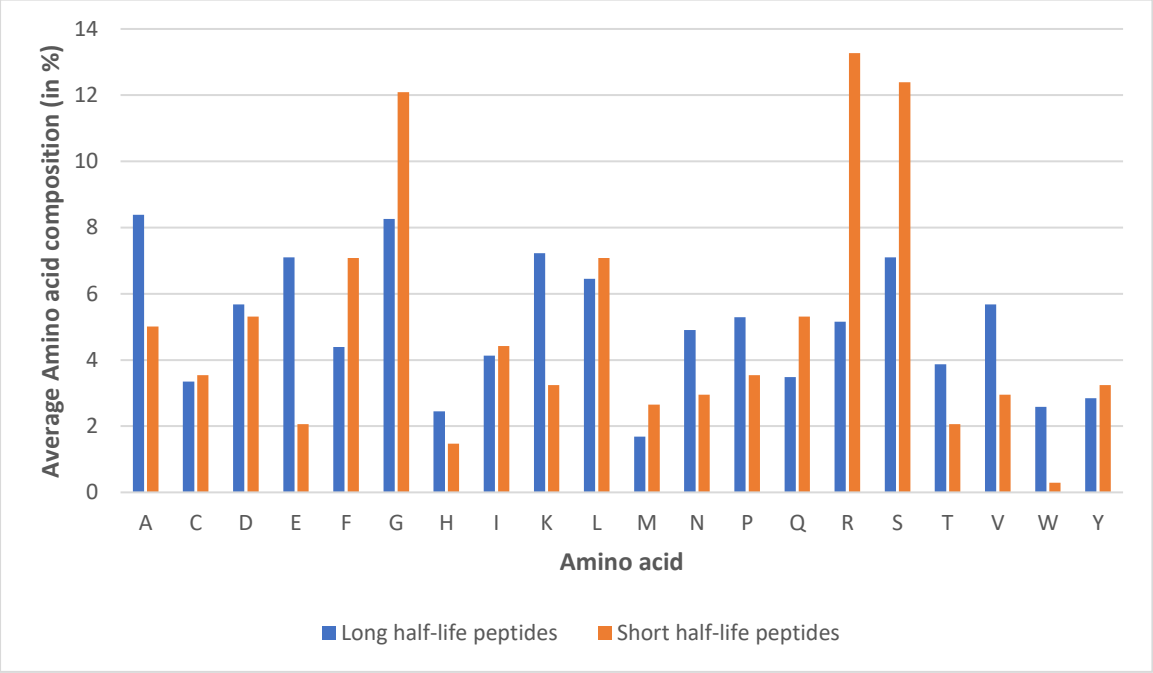


Figure 9: Comparison of amino acid composition in the 20 peptides with the highest and lowest half-lives

The average physico-chemical properties of peptides with long and short half-lives were compared. Peptides with longer half-lives were found to be more aliphatic, hydrophobic, sulfur-rich, negatively charged, and less cyclic than those with shorter half-lives, as shown in Figure 10.

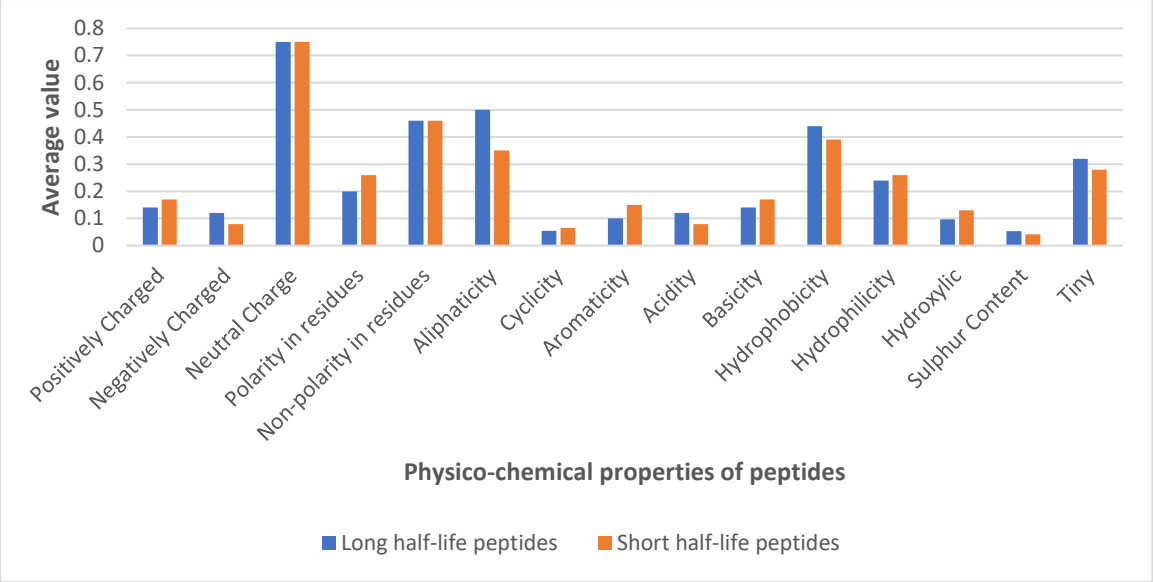


Figure 10: Comparison of Physico-chemical properties in the 20 peptides with the highest and lowest half-lives

To evaluate the impact of various modifications such as N-terminal and C-terminal alterations, incorporation of non-natural amino acids, and integration of D-amino acids on peptide half-life, the top 5 peptides with each modification were compared. The analysis revealed that N-terminal and C-terminal modifications had the most significant effect on extending half-life, while non-natural amino acids had a moderate impact, and D-amino acid incorporation showed only a minimal effect as shown in Table 4.

Table 4: Top 5 peptide half-lives (seconds) in natural form, with N-terminal and C-terminal modifications, and containing non-natural or D-amino acids.

Natural	N-terminal	C-terminal	Non-natural	D-amino acid
831600	4700160	5580000	874800	86400
604800	2419200	1296000	609984	42480
604800	475200	843120	471600	32400
367200	421200	831600	374400	25200
288000	378000	727200	272916	21600

Prediction of peptide half-life using various Machine learning model:

Material and Methods:

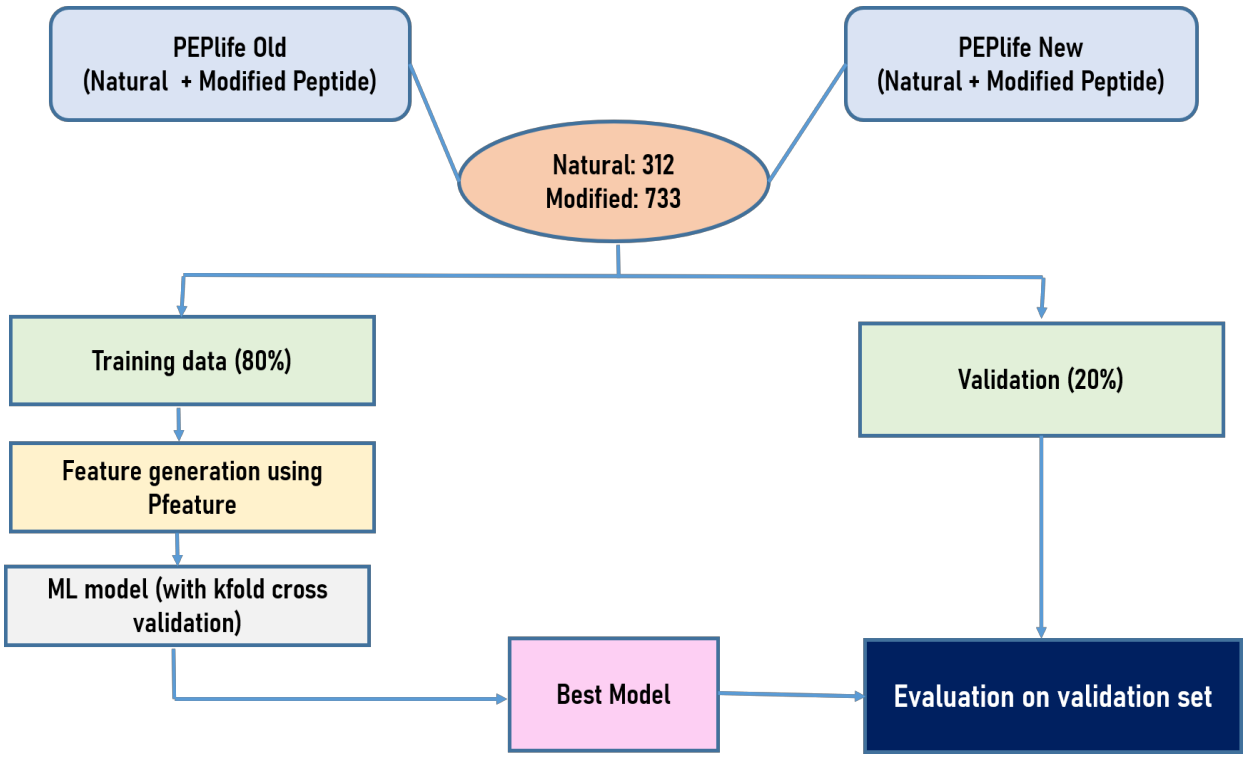


Figure 11: Structural overview of the predictive model for peptide half-life

Most features were generated using **Pfeature**, but for modified peptides, additional binary profiles were included assigning a value of **1** if the peptide was N-terminally modified, C-terminally modified, contained D-amino acids, or was cyclic otherwise, a value of **0** was used. Also, dataset used for Machine learning were peptides having length ranges from 5 to 50.

RESULTS AND DISCUSSION:

Performance of prediction model:

For Natural peptides:

Various machine learning models were trained separately using individual feature sets such as AAC, DPC, and ATC. Initially, models like SVR, Linear Regression, and Ridge were evaluated using 20 AAC features. Among them, **XGBoost** demonstrated the best performance, achieving a

training R^2 of 0.66 with an RMSE of 1.31, and a validation R^2 of 0.29 with an RMSE of 1.82 as shown in Table 5.

Table 5: ML performance for Natural peptide using AAC composition

Amino acid composition (AAC)										
	Training					Validation				
Model	**MSE**	**MAE**	**RMSE**	**PCC**	**R2**	**MSE**	**MAE**	**RMSE**	**PCC**	**R2**
SVR	2.89	1.17	1.7	0.69	0.42	3.63	1.49	1.91	0.56	0.22
Linear Regression	3.66	1.47	1.91	0.53	0.27	4.72	1.6	2.17	0.32	-0.01
Ridge	3.74	1.48	1.93	0.52	0.26	4.92	1.62	2.22	0.3	-0.06
Lasso	5.12	1.77	2.26		-0.02	4.74	1.78	2.18		-0.02
Random Forest	0.71	0.59	0.84	0.95	0.86	3.49	1.41	1.87	0.61	0.25
Gradient Boosting	0.58	0.55	0.76	0.96	0.88	3.49	1.44	1.87	0.57	0.25
AdaBoost	2.22	1.23	1.49	0.8	0.56	4.32	1.65	2.08	0.4	0.07
K-Nearest Neighbors	3.21	1.27	1.79	0.61	0.36	3.98	1.49	2	0.47	0.14
XGBoost	1.73	0.98	1.31	0.94	0.66	3.3	1.39	1.82	0.61	0.29
LightGBM	3.27	1.38	1.81	0.71	0.35	3.85	1.58	1.96	0.48	0.17

With DPC features (400), **Random Forest Regressor** demonstrated the best performance, achieving a training R^2 of 0.79 with an RMSE of 1.04, and a validation R^2 of 0.3 with an RMSE of 1.8 as shown in Table 6.

Table 6: ML performance for Natural peptide using DPC composition

Dipeptide Composition (DPC)										
	Training					Validation				
Model	**MSE**	**MAE**	**RMSE**	**PCC**	**R2**	**MSE**	**MAE**	**RMSE**	**PCC**	**R2**
SVR	2.63	1.12	1.62	0.73	0.48	3.3	1.44	1.82	0.67	0.29
Linear Regression	1.68	0.96	1.29	0.82	0.67	7.36	1.82	2.71	0.37	-0.58
Ridge	1.67	0.96	1.29	0.82	0.67	6.85	1.78	2.62	0.38	-0.47
Lasso	5.12	1.77	2.26		-0.02	4.74	1.78	2.18		-0.02
Random Forest	1.08	0.76	1.04	0.91	0.79	3.25	1.3	1.8	0.6	0.3
Gradient Boosting	1.08	0.79	1.04	0.91	0.78	3.91	1.43	1.98	0.49	0.16
AdaBoost	3.79	1.5	1.94	0.57	0.25	4.49	1.69	2.12	0.56	0.04
K-Nearest Neighbors	2.64	1.21	1.62	0.7	0.47	3.57	1.46	1.89	0.52	0.23
XGBoost	1.2	0.8	1.09	0.9	0.76	3.44	1.37	1.85	0.58	0.26
LightGBM	3.63	1.45	1.9	0.57	0.28	3.86	1.56	1.97	0.58	0.17

With ATC features (4), **KNN** demonstrated the best performance, achieving a training R^2 of 0.36 with an RMSE of 1.79, and a validation R^2 of 0.3 with an RMSE of 1.81 as shown in Table 7.

Table 7: ML performance for Natural peptide using ATC composition

Atomic Composition (ATC)										
	Training					Validation				
Model	**MSE**	**MAE**	**RMSE**	**PCC**	**R2**	**MSE**	**MAE**	**RMSE**	**PCC**	**R2**
SVR	4.09	1.5	2.02	0.47	0.19	4.45	1.69	2.11	0.29	0.04
Linear Regression	4.93	1.72	2.22	0.19	0.02	4.96	1.83	2.23	0	-0.06
Ridge	5.03	1.74	2.24	0.13	0	4.8	1.82	2.19	0.08	-0.03
Lasso	5.12	1.77	2.26		-0.02	4.74	1.78	2.18		-0.02
Random Forest	0.75	0.63	0.87	0.96	0.85	4.67	1.68	2.16	0.32	0
Gradient Boosting	0.89	0.69	0.94	0.94	0.82	5.42	1.8	2.33	0.21	-0.17
AdaBoost	2.98	1.41	1.72	0.71	0.41	5.56	1.92	2.36	0.09	-0.2
K-Nearest Neighbors	3.21	1.37	1.79	0.6	0.36	3.27	1.41	1.81	0.59	0.3
XGBoost	2.27	1.14	1.51	0.89	0.55	4.48	1.64	2.12	0.29	0.04
LightGBM	3.85	1.53	1.96	0.61	0.23	4.52	1.74	2.13	0.26	0.03

For modified peptides:

With AAC features (20), **LightGBM** demonstrated the best performance, achieving a training R^2 of 0.99 with an RMSE of 0.5, and a validation R^2 of 0.67 with an RMSE of 2.67 as shown in Table 8.

Table 8: ML performance for Modified peptide using AAC composition

Amino acid composition (AAC)										
	Training					Validation				
Model	**MSE**	**MAE**	**RMSE**	**PCC**	**R2**	**MSE**	**MAE**	**RMSE**	**PCC**	**R2**
SVR	0.73	0.53	0.85	0.98	0.97	8.68	1.77	2.95	0.8	0.61
Linear Regression	19.1	3	4.37	0.52	0.09	39.66	3.5	6.3	0.4	-0.8
Ridge	19.2	3.01	4.38	0.52	0.08	40.68	3.51	6.38	0.4	-0.85
Lasso	25.67	4.23	5.07		-0.22	27.21	4.34	5.22		-0.24
Random Forest	1.98	0.82	1.4	0.96	0.91	9.1	1.76	3.02	0.8	0.59
Gradient Boosting	2.41	0.95	1.55	0.95	0.89	9.31	1.87	3.05	0.78	0.58
AdaBoost	8.14	2.1	2.85	0.91	0.61	13.48	2.61	3.67	0.8	0.39
K-Nearest Neighbors	5.58	1.33	2.36	0.87	0.73	8.8	1.67	2.97	0.8	0.6
XGBoost	0	0.02	0.06	1	1	8.81	1.65	2.97	0.79	0.6
LightGBM	0.26	0.28	0.5	0.99	0.99	7.15	1.55	2.67	0.83	0.67

With DPC features (400), **KNN** demonstrated the best performance, achieving a training R^2 of 0.85 with an RMSE of 1.78, and a validation R^2 of 0.83 with an RMSE of 1.83 as shown in Table 9.

Table 9: ML performance for Modified peptide using DPC composition

Dipeptide Composition (DPC)										
	Training					Validation				
Model	**MSE**	**MAE**	**RMSE**	**PCC**	**R2**	**MSE**	**MAE**	**RMSE**	**PCC**	**R2**
SVR	0.56	0.19	0.75	0.99	0.97	6.31	1.48	2.51	0.84	0.69
Linear Regression	4.99	1.37	2.23	0.89	0.77	36.11	3	6.01	0.65	-0.78
Ridge	4.97	1.37	2.23	0.89	0.77	35.05	2.97	5.92	0.65	-0.73
Lasso	26.2	4.25	5.12		-0.22	25.24	4.26	5.02		-0.24
Random Forest	0.99	0.52	0.99	0.98	0.95	4.1	1.21	2.02	0.9	0.8
Gradient Boosting	4.84	1.42	2.2	0.92	0.77	9.42	2.05	3.07	0.76	0.54
AdaBoost	18.29	3.19	4.27	0.57	0.15	18.13	3.43	4.26	0.63	0.11
K-Nearest Neighbors	3.18	1.02	1.78	0.93	0.85	3.35	1.04	1.83	0.92	0.83
XGBoost	0.41	0.24	0.64	0.99	0.98	8.51	1.74	2.92	0.8	0.58
LightGBM	3.04	1.04	1.74	0.93	0.86	6.35	1.57	2.52	0.84	0.69

With ATC features (4), **Ranom Forest Regressor** demonstrated the best performance, achieving a training R^2 of 0.82 with an RMSE of 1.95, and a validation R^2 of 0.3 with an RMSE of 3.76 as shown in Table 10.

Table 10: ML performance for Modified peptide using ATC composition

Atomic Composition (ATC)										
	Training					Validation				
Model	**MSE**	**MAE**	**RMSE**	**PCC**	**R2**	**MSE**	**MAE**	**RMSE**	**PCC**	**R2**
SVR	14.34	2.61	3.79	0.62	0.33	15.16	2.75	3.89	0.57	0.25
Linear Regression	22.11	3.6	4.7	0.36	-0.03	24.3	3.78	4.93	0.25	-0.2
Ridge	22.42	3.63	4.73	0.35	-0.05	24.15	3.77	4.91	0.26	-0.19
Lasso	26.2	4.25	5.12		-0.22	25.24	4.26	5.02		-0.24
Random Forest	3.8	1.18	1.95	0.94	0.82	14.15	2.57	3.76	0.62	0.3
Gradient Boosting	8.46	1.91	2.91	0.82	0.6	15.65	2.87	3.96	0.56	0.23
AdaBoost	18.83	3.38	4.34	0.59	0.12	20.19	3.52	4.49	0.4	0
K-Nearest Neighbors	12.67	2.39	3.56	0.69	0.41	14.9	2.65	3.86	0.59	0.27
XGBoost	3.74	1.11	1.93	0.93	0.83	15.48	2.74	3.93	0.57	0.24
LightGBM	7.48	1.75	2.73	0.83	0.65	14.96	2.74	3.87	0.59	0.26

The best-performing model for **modified peptides** was **K-Nearest Neighbors (KNN)** using **Dipeptide Composition (DPC)** features, combined with binary indicators for N-terminal modification, C-terminal modification, presence of D-amino acids, and non-natural residues. This model achieved a **training R^2 of 0.85** with an **RMSE of 1.78**, and a **validation R^2 of 0.83** with an **RMSE of 1.83**, indicating strong performance without signs of overfitting.

In contrast, for **natural peptides**, model performance was lower. The best result came from **XGBoost** with **Amino Acid Composition (AAC)** features, reaching a **training R^2 of 0.66** (RMSE 1.31) and a **validation R^2 of only 0.29** (RMSE 1.82).

This suggests that more advanced approaches, such as **deep learning** or **large language models (LLMs)**, could be explored in the future to improve prediction accuracy, especially for natural peptides.

Chapter-4

Conclusion

Peptides are promising therapeutic agents due to their high specificity and broad physiological roles, including functioning as hormones, neurotransmitters, and antimicrobial agents. However, their clinical development faces challenges such as enzymatic degradation, poor oral bioavailability, and rapid clearance. To improve stability and efficacy, several modification strategies have been explored, including cyclization, incorporation of D-amino acids and unnatural residues, and advanced delivery systems.

To support rational peptide design, we developed **PEPlife2** (https://webs.iiitd.edu.in/raghava/peplife2/), a significantly expanded database containing **4,486 entries** with detailed annotations on sequence features, modifications, biological activity, assay conditions, and half-life data. PEPlife2 introduces tools like 3D visualization, residue-level modification mapping, and an API for programmatic access.

We also implemented machine learning models to predict peptide half-life. The best-performing model for **modified peptides** was **KNN using dipeptide composition (DPC)** and binary indicators for structural modifications, achieving a **training R^2 of 0.85** and **validation R^2 of 0.83**. For **natural peptides**, performance was lower, with **XGBoost using amino acid composition (AAC)** achieving a **training R^2 of 0.66** and **validation R^2 of 0.29**. These results suggest that more advanced models, such as **deep learning** or **large language models (LLMs)**, could further improve prediction accuracy.

Overall, PEPlife2 is a valuable resource for peptide research, advancing the development of more stable and effective peptide-based therapeutics.

.

Bibliography

1. Henninot A, Collins JC, Nuss JM. The Current State of Peptide Drug Discovery: Back to the Future? J Med Chem. 2018 Feb 22;61(4):1382-1414. doi: 10.1021/acs.jmedchem.7b00318. Epub 2017 Aug 11. PMID: 28737935.Fosgerau K, Hoffmann T. Peptide therapeutics: current status and future directions. Drug Discov Today. 2015 Jan;20(1):122-8. doi: 10.1016/j.drudis.2014.10.003. Epub 2014 Oct 17. PMID: 25450771.
2. Giordano C, Marchiò M, Timofeeva E, Biagini G. Neuroactive peptides as putative mediators of antiepileptic ketogenic diets. Front Neurol. 2014 Apr 29;5:63. doi: 10.3389/fneur.2014.00063. PMID: 24808888; PMCID: PMC4010764.
3. Davda J, Declerck P, Hu-Lieskovan S, Hickling TP, Jacobs IA, Chou J, Salek-Ardakani S, Kraynov E. Immunogenicity of immunomodulatory, antibody-based, oncology therapeutics. J Immunother Cancer. 2019 Apr 15;7(1):105. doi: 10.1186/s40425-019-0586-0. PMID: 30992085; PMCID: PMC6466770.
4. Waldmann H. Human monoclonal antibodies: the residual challenge of antibody immunogenicity. Methods Mol Biol. 2014;1060:1-8. doi: 10.1007/978-1-62703-586-6_1. PMID: 24037833.
5. Fosgerau K, Hoffmann T. Peptide therapeutics: current status and future directions. Drug Discov Today. 2015 Jan;20(1):122-8. doi: 10.1016/j.drudis.2014.10.003. Epub 2014 Oct 17. PMID: 25450771.
6. Muttenthaler M, King GF, Adams DJ, Alewood PF. Trends in peptide drug discovery. Nat Rev Drug Discov. 2021 Apr;20(4):309-325. doi: 10.1038/s41573-020-00135-8. Epub 2021 Feb 3. PMID: 33536635.
7. Furusawa S, Nomoto H, Yokoyama H, Suzuki Y, Tsuzuki A, Takahashi K, Miya A, Kameda H, Cho KY, Takeuchi J, Nagai S, Taneda S, Kurihara Y, Nakamura A, Atsumi T; SWITCH-SEMA 2 study group. Glycaemic control efficacy of switching from dipeptidyl peptidase-4 inhibitors to oral semaglutide in subjects with type 2 diabetes: A multicentre, prospective, randomized, open-label, parallel-group comparison study (SWITCH-SEMA 2 study). Diabetes Obes Metab. 2024 Mar;26(3):961-970. doi: 10.1111/dom.15393. Epub 2023 Dec 11. PMID: 38073422.
8. Qin W, Yang J, Deng C, Ruan Q, Duan K. Efficacy and safety of semaglutide 2.4 mg for weight loss in overweight or obese adults without diabetes: An updated systematic review and meta-analysis including the 2-year STEP 5 trial. Diabetes Obes Metab. 2024 Mar;26(3):911-923. doi: 10.1111/dom.15386. Epub 2023 Nov 28. PMID: 38016699.
9. Xiao W, Jiang W, Chen Z, Huang Y, Mao J, Zheng W, Hu Y, Shi J. Advance in peptide-based drug development: delivery platforms, therapeutics and vaccines. Signal Transduct

Target Ther. 2025 Mar 5;10(1):74. doi: 10.1038/s41392-024-02107-5. PMID: 40038239; PMCID: PMC11880366.

10. Trujillo JM, Nuffer W. GLP-1 receptor agonists for type 2 diabetes mellitus: recent developments and emerging agents. Pharmacotherapy. 2014 Nov;34(11):1174-86. doi: 10.1002/phar.1507. PMID: 25382096.
11. Wang L, Wang N, Zhang W, Cheng X, Yan Z, Shao G, Wang X, Wang R, Fu C. Therapeutic peptides: current applications and future directions. Signal Transduct Target Ther. 2022 Feb 14;7(1):48. doi: 10.1038/s41392-022-00904-4. PMID: 35165272; PMCID: PMC8844085.
12. Lee AC, Harris JL, Khanna KK, Hong JH. A Comprehensive Review on Current Advances in Peptide Drug Development and Design. Int J Mol Sci. 2019 May 14;20(10):2383. doi: 10.3390/ijms20102383. PMID: 31091705; PMCID: PMC6566176.
13. Fernandez Rico C, Konate K, Josse E, Nargeot J, Barrère-Lemaire S, Boisguérin P. Therapeutic Peptides to Treat Myocardial Ischemia-Reperfusion Injury. Front Cardiovasc Med. 2022 Feb 17;9:792885. doi: 10.3389/fcvm.2022.792885. PMID: 35252383; PMCID: PMC8891520.
14. Wang H, Zhai F. Programme and policy options for preventing obesity in China. Obes Rev. 2013 Nov;14 Suppl 2(0 2):134-40. doi: 10.1111/obr.12106. PMID: 24102781; PMCID: PMC4048452.
15. Jin J. JAMA PATIENT PAGE. Specific Medications for Weight Loss. JAMA. 2015 Aug 18;314(7):742. doi: 10.1001/jama.2015.5836. PMID: 26284734.
16. Ahrens VM, Bellmann-Sickert K, Beck-Sickinger AG. Peptides and peptide conjugates: therapeutics on the upward path. Future Med Chem. 2012 Aug;4(12):1567-86. doi: 10.4155/fmc.12.76. PMID: 22917246.
17. Dodd GT, Decherf S, Loh K, Simonds SE, Wiede F, Balland E, Merry TL, Münzberg H, Zhang ZY, Kahn BB, Neel BG, Bence KK, Andrews ZB, Cowley MA, Tiganis T. Leptin and insulin act on POMC neurons to promote the browning of white fat. Cell. 2015 Jan 15;160(1-2):88-104. doi: 10.1016/j.cell.2014.12.022. PMID: 25594176; PMCID: PMC4453004.
18. Pandey SC. Anxiety and alcohol abuse disorders: a common role for CREB and its target, the neuropeptide Y gene. Trends Pharmacol Sci. 2003 Sep;24(9):456-60. doi: 10.1016/S0165-6147(03)00226-8. PMID: 12967770.
19. Gao Y, Yuan X, Zhu Z, Wang D, Liu Q, Gu W. Research and prospect of peptides for use in obesity treatment (Review). Exp Ther Med. 2020 Dec;20(6):234. doi: 10.3892/etm.2020.9364. Epub 2020 Oct 16. PMID: 33149788; PMCID: PMC7604735.

20. Tyagi A, Tuknait A, Anand P, Gupta S, Sharma M, Mathur D, Joshi A, Singh S, Gautam A, Raghava GP. CancerPPD: a database of anticancer peptides and proteins. Nucleic Acids Res. 2015 Jan;43(Database issue):D837-43. doi: 10.1093/nar/gku892. Epub 2014 Sep 30. PMID: 25270878; PMCID: PMC4384006.
21. Farokhzad OC, Langer R. Impact of nanotechnology on drug delivery. ACS Nano. 2009 Jan 27;3(1):16-20. doi: 10.1021/nn900002m. PMID: 19206243.
22. Garay H, Espinosa LA, Perera Y, Sánchez A, Diago D, Perea SE, Besada V, Reyes O, González LJ. Characterization of low-abundance species in the active pharmaceutical ingredient of CIGB-300: A clinical-grade anticancer synthetic peptide. J Pept Sci. 2018 Jun;24(6):e3081. doi: 10.1002/psc.3081. Epub 2018 Apr 20. PMID: 29676523.
23. Perea SE, Reyes O, Baladron I, Perera Y, Farina H, Gil J, Rodriguez A, Bacardi D, Marcelo JL, Cosme K, Cruz M, Valenzuela C, López-Saura PA, Puchades Y, Serrano JM, Mendoza O, Castellanos L, Sanchez A, Betancourt L, Besada V, Silva R, López E, Falcón V, Hernández I, Solares M, Santana A, Díaz A, Ramos T, López C, Ariosa J, González LJ, Garay H, Gómez D, Gómez R, Alonso DF, Sigman H, Herrera L, Acevedo B. CIGB-300, a novel proapoptotic peptide that impairs the CK2 phosphorylation and exhibits anticancer properties both in vitro and in vivo. Mol Cell Biochem. 2008 Sep;316(1-2):163-7. doi: 10.1007/s11010-008-9814-5. Epub 2008 Jun 25. PMID: 18575815.
24. Rodríguez-Ulloa A, Ramos Y, Gil J, Perera Y, Castellanos-Serra L, García Y, Betancourt L, Besada V, González LJ, Fernández-de-Cossio J, Sanchez A, Serrano JM, Farina H, Alonso DF, Acevedo BE, Padrón G, Musacchio A, Perea SE. Proteomic profile regulated by the anticancer peptide CIGB-300 in non-small cell lung cancer (NSCLC) cells. J Proteome Res. 2010 Oct 1;9(10):5473-83. doi: 10.1021/pr100728v. PMID: 20804217.
25. Yeaman MR, Yount NY. Mechanisms of antimicrobial peptide action and resistance. Pharmacol Rev. 2003 Mar;55(1):27-55. doi: 10.1124/pr.55.1.2. PMID: 12615953.
26. Hancock RE, Sahl HG. Antimicrobial and host-defense peptides as new anti-infective therapeutic strategies. Nat Biotechnol. 2006 Dec;24(12):1551-7. doi: 10.1038/nbt1267. PMID: 17160061.
27. Pasupuleti M, Schmidtchen A, Malmsten M. Antimicrobial peptides: key components of the innate immune system. Crit Rev Biotechnol. 2012 Jun;32(2):143-71. doi: 10.3109/07388551.2011.594423. Epub 2011 Nov 11. PMID: 22074402.
28. Lamb HM, Wiseman LR. Pexiganan acetate. Drugs. 1998 Dec;56(6):1047-52; discussion 1053-4. doi: 10.2165/00003495-199856060-00011. PMID: 9878992.
29. Lipsky BA, Holroyd KJ, Zasloff M. Topical versus systemic antimicrobial therapy for treating mildly infected diabetic foot ulcers: a randomized, controlled, double-blinded,

multicenter trial of pexiganan cream. Clin Infect Dis. 2008 Dec 15;47(12):1537-45. doi: 10.1086/593185. PMID: 18990064.

30. Clarfield AM. The decreasing prevalence of reversible dementias: an updated meta-analysis. Arch Intern Med. 2003 Oct 13;163(18):2219-29. doi: 10.1001/archinte.163.18.2219. PMID: 14557220.
31. Dutta R, Trapp BD. Mechanisms of neuronal dysfunction and degeneration in multiple sclerosis. Prog Neurobiol. 2011 Jan;93(1):1-12. doi: 10.1016/j.pneurobio.2010.09.005. Epub 2010 Oct 12. PMID: 20946934; PMCID: PMC3030928.
32. Pardridge WM. Blood-Brain Barrier and Delivery of Protein and Gene Therapeutics to Brain. Front Aging Neurosci. 2020 Jan 10;11:373. doi: 10.3389/fnagi.2019.00373. PMID: 31998120; PMCID: PMC6966240.
33. Coyle J, Kershaw P. Galantamine, a cholinesterase inhibitor that allosterically modulates nicotinic receptors: effects on the course of Alzheimer's disease. Biol Psychiatry. 2001 Feb 1;49(3):289-99. doi: 10.1016/s0006-3223(00)01101-x. PMID: 11230880.
34. Byerley WF, Reimherr FW, Wood DR, Grosser BI. Fluoxetine, a selective serotonin uptake inhibitor, for the treatment of outpatients with major depression. J Clin Psychopharmacol. 1988 Apr;8(2):112-5. PMID: 3286684.
35. Puente XS, Gutiérrez-Fernández A, Ordóñez GR, Hillier LW, López-Otín C. Comparative genomic analysis of human and chimpanzee proteases. Genomics. 2005 Dec;86(6):638-47. doi: 10.1016/j.ygeno.2005.07.009. Epub 2005 Sep 12. PMID: 16162398.
36. Di L. Strategic approaches to optimizing peptide ADME properties. AAPS J. 2015 Jan;17(1):134-43. doi: 10.1208/s12248-014-9687-3. Epub 2014 Nov 4. PMID: 25366889; PMCID: PMC4287298.
37. Kataoka M, Fukahori M, Ikemura A, Kubota A, Higashino H, Sakuma S, Yamashita S. Effects of gastric pH on oral drug absorption: In vitro assessment using a dissolution/permeation system reflecting the gastric dissolution process. Eur J Pharm Biopharm. 2016 Apr;101:103-11. doi: 10.1016/j.ejpb.2016.02.002. Epub 2016 Feb 9. PMID: 26873006.
38. Verma S, Goand UK, Husain A, Katekar RA, Garg R, Gayen JR. Challenges of peptide and protein drug delivery by oral route: Current strategies to improve the bioavailability. Drug Dev Res. 2021 Nov;82(7):927-944. doi: 10.1002/ddr.21832. Epub 2021 May 14. PMID: 33988872.
39. Werle M, Bernkop-Schnürch A. Strategies to improve plasma half life time of peptide and protein drugs. Amino Acids. 2006 Jun;30(4):351-67. doi: 10.1007/s00726-005-0289-3. Epub 2006 Apr 20. PMID: 16622600.

40. Al Musaimi O, Lombardi L, Williams DR, Albericio F. Strategies for Improving Peptide Stability and Delivery. Pharmaceuticals (Basel). 2022 Oct 19;15(10):1283. doi: 10.3390/ph15101283. PMID: 36297395; PMCID: PMC9610364.
41. Derakhshankhah H, Jafari S. Cell penetrating peptides: A concise review with emphasis on biomedical applications. Biomed Pharmacother. 2018 Dec;108:1090-1096. doi: 10.1016/j.biopha.2018.09.097. Epub 2018 Sep 28. PMID: 30372809.
42. Stiltner J, McCandless K, Zahid M. Cell-Penetrating Peptides: Applications in Tumor Diagnosis and Therapeutics. Pharmaceutics. 2021 Jun 15;13(6):890. doi: 10.3390/pharmaceutics13060890. PMID: 34204007; PMCID: PMC8232808.
43. Khan MM, Filipczak N, Torchilin VP. Cell penetrating peptides: A versatile vector for co-delivery of drug and genes in cancer. J Control Release. 2021 Feb 10;330:1220-1228. doi: 10.1016/j.jconrel.2020.11.028. Epub 2020 Nov 25. PMID: 33248708.
44. Desale K, Kuche K, Jain S. Cell-penetrating peptides (CPPs): an overview of applications for improving the potential of nanotherapeutics. Biomater Sci. 2021 Feb 23;9(4):1153-1188. doi: 10.1039/d0bm01755h. PMID: 33355322.
45. Xie J, Bi Y, Zhang H, Dong S, Teng L, Lee RJ, Yang Z. Cell-Penetrating Peptides in Diagnosis and Treatment of Human Diseases: From Preclinical Research to Clinical Application. Front Pharmacol. 2020 May 20;11:697. doi: 10.3389/fphar.2020.00697. PMID: 32508641; PMCID: PMC7251059.
46. Verdine GL, Walensky LD. The challenge of drugging undruggable targets in cancer: lessons learned from targeting BCL-2 family members. Clin Cancer Res. 2007 Dec 15;13(24):7264-70. doi: 10.1158/1078-0432.CCR-07-2184. PMID: 18094406.
47. Zorzi A, Deyle K, Heinis C. Cyclic peptide therapeutics: past, present and future. Curr Opin Chem Biol. 2017 Jun;38:24-29. doi: 10.1016/j.cbpa.2017.02.006. Epub 2017 Feb 27. PMID: 28249193.
48. Jamieson AG, Boutard N, Sabatino D, Lubell WD. Peptide scanning for studying structure-activity relationships in drug discovery. Chem Biol Drug Des. 2013 Jan;81(1):148-65. doi: 10.1111/cbdd.12042. PMID: 23253136.
49. Domhan C, Uhl P, Kleist C, Zimmermann S, Umstätter F, Leotta K, Mier W, Wink M. Replacement of l-Amino Acids by d-Amino Acids in the Antimicrobial Peptide Ranalexin and Its Consequences for Antimicrobial Activity and Biodistribution. Molecules. 2019 Aug 17;24(16):2987. doi: 10.3390/molecules24162987. PMID: 31426494; PMCID: PMC6720431.
50. Eckhauser ML. The neodymium-YAG laser and gastrointestinal malignancy. Arch Surg. 1990 Sep;125(9):1152-4. doi: 10.1001/archsurg.1990.01410210078012. PMID: 1698048.

51. Li Y, Li W, Xu Z. Improvement on Permeability of Cyclic Peptide/Peptidomimetic: Backbone *N*-Methylation as A Useful Tool. Mar Drugs. 2021 May 27;19(6):311. doi: 10.3390/md19060311. PMID: 34072121; PMCID: PMC8229464.
52. Acar H, Ting JM, Srivastava S, LaBelle JL, Tirrell MV. Molecular engineering solutions for therapeutic peptide delivery. Chem Soc Rev. 2017 Oct 30;46(21):6553-6569. doi: 10.1039/c7cs00536a. PMID: 28902203.
53. Borro BC, Malmsten M. Complexation between antimicrobial peptides and polyelectrolytes. Adv Colloid Interface Sci. 2019 Aug;270:251-260. doi: 10.1016/j.cis.2019.07.001. Epub 2019 Jul 4. PMID: 31301601.
54. Doostmohammadi M, Ameri A, Mohammadinejad R, Dehghannoudeh N, Banat IM, Ohadi M, Dehghannoudeh G. Hydrogels For Peptide Hormones Delivery: Therapeutic And Tissue Engineering Applications. Drug Des Devel Ther. 2019 Sep 26;13:3405-3418. doi: 10.2147/DDDT.S217211. PMID: 31579238; PMCID: PMC6770672.
55. Du AW, Stenzel MH. Drug carriers for the delivery of therapeutic peptides. Biomacromolecules. 2014 Apr 14;15(4):1097-114. doi: 10.1021/bm500169p. Epub 2014 Mar 24. PMID: 24661025.
56. Saleh T, Shojaosadati SA. Multifunctional nanoparticles for cancer immunotherapy. Hum Vaccin Immunother. 2016 Jul 2;12(7):1863-75. doi: 10.1080/21645515.2016.1147635. Epub 2016 Feb 22. PMID: 26901287; PMCID: PMC4964832.
57. Ladner RC, Sato AK, Gorzelany J, de Souza M. Phage display-derived peptides as therapeutic alternatives to antibodies. Drug Discov Today. 2004 Jun 15;9(12):525-9. doi: 10.1016/S1359-6446(04)03104-6. PMID: 15183160.
58. Sato AK, Viswanathan M, Kent RB, Wood CR. Therapeutic peptides: technological advances driving peptides into development. Curr Opin Biotechnol. 2006 Dec;17(6):638-42. doi: 10.1016/j.copbio.2006.10.002. Epub 2006 Oct 17. PMID: 17049837.
59. Adessi C, Soto C. Converting a peptide into a drug: strategies to improve stability and bioavailability. Curr Med Chem. 2002 May;9(9):963-78. doi: 10.2174/0929867024606731. PMID: 11966456.
60. Werle M, Bernkop-Schnürch A. Strategies to improve plasma half life time of peptide and protein drugs. Amino Acids. 2006 Jun;30(4):351-67. doi: 10.1007/s00726-005-0289-3. Epub 2006 Apr 20. PMID: 16622600.
61. Holz E, Darwish M, Tesar DB, Shatz-Binder W. A Review of Protein- and Peptide-Based Chemical Conjugates: Past, Present, and Future. Pharmaceutics. 2023 Feb 10;15(2):600. doi: 10.3390/pharmaceutics15020600. PMID: 36839922; PMCID: PMC9959917.

62. Mathur D, Prakash S, Anand P, Kaur H, Agrawal P, Mehta A, Kumar R, Singh S, Raghava GP. PEPlife: A Repository of the Half-life of Peptides. Sci Rep. 2016 Nov 7;6:36617. doi: 10.1038/srep36617. PMID: 27819351; PMCID: PMC5098197.
63. D'Aloisio V, Dognini P, Hutcheon GA, Coxon CR. PepTherDia: database and structural composition analysis of approved peptide therapeutics and diagnostics. Drug Discov Today. 2021 Jun;26(6):1409-1419. doi: 10.1016/j.drudis.2021.02.019. Epub 2021 Feb 26. PMID: 33647438.
64. Mathur D, Singh S, Mehta A, Agrawal P, Raghava GPS. In silico approaches for predicting the half-life of natural and modified peptides in blood. PLoS One. 2018 Jun 1;13(6):e0196829. doi: 10.1371/journal.pone.0196829. PMID: 29856745; PMCID: PMC5983457.
65. S. Singh, H. Singh, A. Tuknait, K. Chaudhary, B. Singh, S. Kumaran, G.P.S. Raghava, PEPstrMOD: structure prediction of peptides containing natural, non-natural and modified residues, Biol. Direct 10 (2015) 73. https://doi.org/10.1186/s13062-015-0103-4.
66. J. Yang, R. Yan, A. Roy, D. Xu, J. Poisson, Y. Zhang, The I-TASSER Suite: protein structure and function prediction, Nat. Methods 12 (2015) 7–8. https://doi.org/10.1038/nmeth.3213.
67. W. Kabsch, C. Sander, Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features, Biopolymers 22 (1983) 2577–2637. https://doi.org/10.1002/bip.360221211.
68. T.F. Smith, M.S. Waterman, Identification of common molecular subsequences, J. Mol. Biol. 147 (1981) 195–197. https://doi.org/10.1016/0022-2836(81)90087-5.
69. S.F. Altschul, W. Gish, W. Miller, E.W. Myers, D.J. Lipman, Basic local alignment search tool, J. Mol. Biol. 215 (1990) 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2.
70. A. Shendre, N.K. Mehta, A.S. Rathore, N. Kumar, S. Patiyal, G.P.S. Raghava, MAP format for representing chemical modifications, annotations, and mutations in protein sequences: An extension of the FASTA format, ArXiv [q-Bio.BM] (2025). https://doi.org/10.48550/ARXIV.2505.03403.