A Comprehensive Compilation of Anticancer Peptides and Prediction of Anticancer Activity in Chemically Modified Peptides

by

Milind Chauhan

Under the Supervision of

Prof. G.P.S Raghava

Indraprastha Institute of Information Technology Delhi

June, 2025

A Comprehensive Compilation of Anticancer Peptides and Prediction of Anticancer Activity in Chemically Modified Peptides

by

Milind Chauhan

Submitted

in partial fulfillment of the requirements for the degree of

Master of Technology

to

Indraprastha Institute of Information Technology Delhi

June, 2025

Certificate

This is to certify that the thesis titled **A Comprehensive Compilation of Anticancer Peptides and Prediction of Anticancer Activity in Chemically Modified Peptides** being submitted by **Milind Chauhan** to the Indraprastha Institute of Information Technology Delhi, for the award of the Master of Technology, is an original research work carried out by him under my supervision. In my opinion, the thesis has reached the standards fulfilling the requirements of the regulations relating to the degree.

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

June, 2025

Prof. G.P.S Raghava
Department of Computational Biology
Indraprastha Institute of Information
Technology Delhi
New Delhi 110020

Acknowledgements

I would like to express my deepest gratitude to all those who have supported me and guided me throughout my M.Tech. thesis work. First and foremost, I would like to thank my esteemed project supervisor **Prof. G.P.S Raghava**, for providing me this opportunity to work under his guidance. His wisdom and guidance have profoundly shaped my understanding and approach to this project.

I am extremely grateful to **Ritu Tomer** for her involvement, insightful comments and motivation. A special mention to my classmate **Amisha Gupta** who has been deeply involved in this project alongside me. Their insights and efforts have been invaluable, and working with them has been a rewarding experience.

Lastly, would like to thank IIIT-Delhi for providing the necessary infrastructure.

Milind Chauhan
MT23232

Abstract

This thesis presents a repository CancerPPD2, which is an updated version of CancerPPD, developed to maintain comprehensive information about Anticancer peptides and proteins. It contains 6521 entries, each entry provides detailed information about an anticancer activity of peptides or proteins that include origin of the peptide, cancer cell line, type of cancer, peptide sequence, and structure. These anticancer peptides have been tested against 392 types of cancer cell lines and 28 types of cancer-associated tissues. In addition to natural anticancer peptides, CancerPPD2 contains 781 entries for chemically modified and 3018 entries for N-/C- terminus modified anticancer peptides. Few entries are also linked with 47 clinical studies and have provided the cross reference to Uniprot and NCT. On average, CancerPPD2 contains around 85% more information than its previous version, CancerPPD. The structures of these anticancer peptides and proteins were either obtained from the Protein Data Bank (PDB) or predicted using PEPstrMOD, I-TASSER, or AlphaFold. A wide range of tools have been integrated into CancerPPD2 for data retrieval and similarity searches. Additionally, we integrated a REST API into this repository to facilitate automatic data retrieval via programs. In this study, we have also made an initial attempt to develop a method for classifying chemically modified anticancer peptides. To the best of our knowledge, no previous efforts have been made in this direction, making this the first study to address this unexplored aspect. This work lays the foundation for future computational models aimed at understanding and predicting the functional behaviour of modified anticancer peptides.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Cancer is a group of diseases characterized by the uncontrolled division of abnormal cells in the human body resulting in catastrophic destruction of normal tissues in the body. Cancer cells proliferate excessively, invade surrounding tissues and may metastasize to distant sites through connective tissues like blood and lymphatic system [1]. It is one of the leading causes of death worldwide. According to ICMR, over 7 Lakh new cancer patients are registered in India every year. Present strategies to combat cancer include chemotherapy, radiation, surgery, or their combinations. Despite extensive research and clinical efforts over the years, conventional therapeutic approaches have demonstrated only limited efficacy, primarily due to factors such as the development of acquired resistance to chemotherapeutic agents [2]. While these strategies can provide symptomatic relief, they are frequently associated with significant side-effects affecting the quality of life of cancer patients. Researchers are now exploring alternative strategies for treating cancer patients, with anticancer peptides and proteins emerging as a promising alternative to traditional cancer therapies [3].

Peptide - & Protein - based drugs are becoming popular as a new class of drugs combining the high target specificity and biological potency of biologics with favorable pharmacokinetic properties typically associated with small molecules, such as oral bioavailability and effective tissue penetration. Their ability to engage challenging molecular targets with high selectivity, modifiable half-lives, and generally lower toxicity and immunogenicity profiles make them attractive candidates in modern drug development pipelines [4], [5]. With the rapid advancement of molecular biology and peptide research, a growing number of short, bioactive peptides have been identified across a wide spectrum of organisms. These peptides exhibit a diverse range of biological functions, including antibacterial, antifungal, antiviral, anticancer, and immunomodulatory activities [6]. Among them, a distinct class of cationic, low-molecular-weight peptides have gained particular attention for its potent ability to selectively target and kill tumor cells. These are collectively termed anticancer peptides (ACPs).In oncology, extensive research is going on in Anticancer peptides as they are considered to be relatively safe. Small molecules are known to exert indiscriminately effects between normal cells and cancerous cells, exacerbating the immunodeficiency of cancer patients. Unlike conventional chemotherapeutic agents, ACPs can be designed to target specific cancer types, often with reduced off-target toxicity. Additionally, several natural anticancer proteins have shown potent tumor-suppressive activity by modulating cell proliferation, angiogenesis, and apoptosis-related pathways [7]. Many of these Anticancer peptides and proteins have shown promising results in various pre-clinical and clinical trials [8, 9]. Figure 1.1 below shows the mechanism of action of Anticancer peptides.

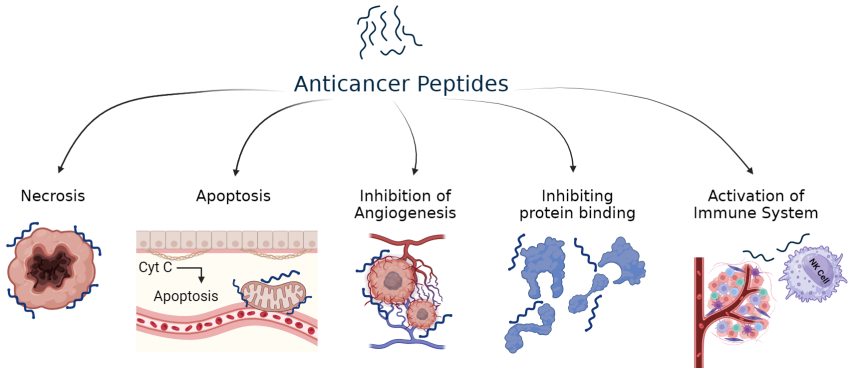


Fig. 1.1. *Mechanism of Action of Anticancer Peptides & Proteins*

Despite their therapeutic promise, information related to ACPs and anticancer proteins is scattered across thousands of publications, patents, and databases, often described in heterogeneous formats and lacking standardization. Experimental data on sequences, structural characteristics, activity profiles, and biological targets is frequently incomplete or inconsistently reported. This fragmentation hampers efforts in data mining, comparative analysis, and rational therapeutic design. Adding further complexity is the growing body of research focused on chemically modified ACPs, which have shown enhanced stability, protease resistance, cell permeability, and bioavailability [8]. These modifications—ranging from N– or C–terminal alterations to non-natural amino acid substitutions and conformational constraints—are crucial for transforming ACPs into viable drug candidates. However, many such peptides are difficult to model computationally, and little to no information is available related to characterization and designing of chemically modified peptides computationally.

Given the accelerating pace of ACP discovery and the increasing interest in chemically engineered variants, there is a compelling need for a centralized, comprehensive, and well-annotated resource that integrates both natural and chemically modified anticancer peptides and proteins. In the following sections, we describe the development of **CancerPPD2.0**, a significantly expanded and updated version of the original CancerPPD database [10]. In addition to this, we present a novel study focused on the prediction of chemically modified anticancer peptides (ACPs), a largely unexplored yet biologically significant area. Recognizing that conventional sequence-based models are often insufficient to capture the structural and chemical complexities introduced by modifications such as D-amino acids, non-natural residues, and terminal modifications, we developed specialized feature representations that integrate both sequence-level and atomic-level information. Using these features, we trained and evaluated a suite of machine learning classifiers to distinguish chemically modified ACPs from non-ACPs.

Chapter 2

Manual Curation of Anticancer Peptides and Proteins

2.1 Introduction

Cancer is among the most life-threatening diseases and represents a significant global challenge in healthcare. Conventional therapies such as chemotherapy are associated with several limitations, including severe side effects, multidrug resistance, non-discriminatory killing of healthy cells and many others. Researchers are continuously exploring novel therapeutic strategies, with peptide- and protein-based approaches emerging as promising alternatives. Extensive research has been conducted on anticancer peptide-based therapeutics; however, the available information remains fragmented and dispersed across the literature. In 2015, Tyagi et al. developed the first such repository that contains extensive information on Anticancer peptides and proteins `CancerPPD` [10]. They manually compiled experimentally validated anticancer peptides and proteins from literature and other public databases. Since the inception of CancerPPD, the number of methods developed for predicting and designing anticancer peptides has grown exponentially. Most existing anticancer peptide prediction methods such as `AntiCP` [11], `ACP-DRL` [12], `MA-PEP` [13], `ACP-ML` [14], `ACPPfel` [15], `AntiCP2` [16], and `MLACP 2.0` [17] have derived datasets from CancerPPD to build their models. This highlights the significance of CancerPPD in cancer biology, particularly in the design of peptides and proteins for treating cancer patients. Since then, nearly a decade later, research in the development of peptide-based anticancer therapeutics has been constantly evolving and growing at a rapid rate. Numerous anticancer peptides have been discovered and experimentally validated by researchers. While these peptides are available in literature, they are not accessible from a single source.

Following the initial effort to establish a comprehensive repository of anticancer peptides and proteins—CancerPPD—in 2015, the past decade has seen an exponential rise in ACP-related research. This surge has led to the discovery of hundreds of novel ACPs and their derivatives. The development of various in silico tools, using data from CancerPPD to classify ACPs, has significantly accelerated this progress. Many of these peptides have demonstrated promising results in preclinical studies, with several formulations advancing to clinical trials. In this study, we present an updated and expanded version of the database, **CancerPPD2**, which offers extensive information on anticancer peptides and proteins. The repository includes cross-references to external resources such as clinical trials status via NCT, structural data from PDB [18] and `AlphaFoldDB` [19], and patent information linked to experimental studies. We believe that CancerPPD2 will serve as a valuable resource for both bioinformatics and experimental researchers working in the field of developing ACP-based therapeutics. Figure 2.1 shows the overall architecture of the database.

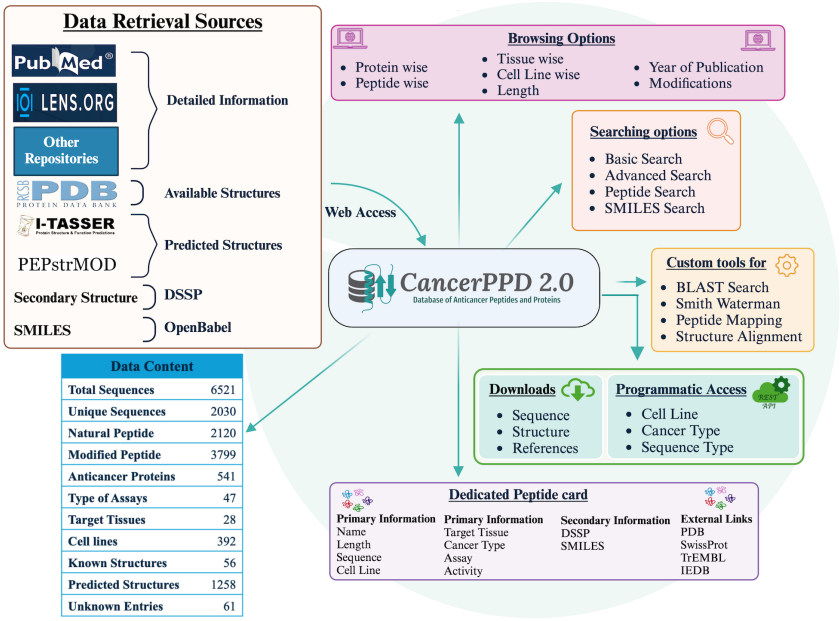


Fig. 2.1. *Overall architecture of CancerPPD2*

2.2 Data Collection & Compilation

To develop a centralized, single-point resource for anticancer peptides and proteins, we systematically collected, curated, and integrated data from diverse scientific and patent literature. We retrieved research articles and patents containing relevant information from multiple databases, including PubMed, Google Scholar, and Patent Lens, focusing on the period from October 2014 to March 2024, as the original CancerPPD database included data only up to 2014, using a combination of keywords like **'ACPs', 'anticancer peptides', 'antitumor peptides', 'anti-angiogenic peptides', 'anti-metastatic peptides' and 'host defense peptides'**. Only experimentally verified ACPs and other relevant experimental information were extracted manually. In addition, information on anticancer proteins was extracted from UniProt, and data related to clinical trials and their current status were retrieved and cross-linked to external resources such as ClinicalTrials.gov (NCT). The database finally comprises **6,521** curated entries, each offering comprehensive information on individual anticancer peptides or proteins, including their origin, peptide sequence, target cancer cell line, cancer type, and structural data. These peptides have been evaluated against **392** distinct cancer cell lines and **28** cancer-associated tissues. In addition to naturally occurring anticancer peptides, the database includes **781** entries for chemically modified peptides and **3,018** entries with N- or C-terminal modifications. Several entries are cross-referenced with **47** clinical studies and linked to external databases such as UniProt. Where available, entries are also associated with clinical trial data. Overall, CancerPPD2 contains approximately 85% more data than its predecessor, CancerPPD. Structural information for these peptides and proteins was either sourced from the Protein

Data Bank (PDB) or predicted using computational tools such as PepstrMOD, I-TASSER, and AlphaFold.

2.3 Database Architecture and Web Interface

The **CancerPPD 2.0** database is deployed on a robust, Linux-based server infrastructure utilizing the Apache HTTP Server for reliable web service delivery. The backend is supported by a MySQL relational database management system (RDBMS), which ensures structured data storage, efficient querying, and transactional integrity of peptide- and protein-related records. To streamline data ingestion, parsing, and archival processes, comprehensive documentation and utility scripts are also provided.

The front-end web interface is fully responsive and cross-platform compatible, developed using modern web technologies such as HTML5, CSS3, JavaScript, and the Bootstrap framework. This design ensures seamless user experiences across desktops, tablets, and mobile devices. Client-side interactivity is achieved through asynchronous JavaScript functionalities, while server-side operations and dynamic content rendering are managed by PHP scripts through the Common Gateway Interface (CGI) protocol. Furthermore, the database architecture is optimized for interoperability and programmatic access. A RESTful API is integrated into the backend, enabling automated querying, bulk data extraction, and seamless integration with external computational workflows or bioinformatics pipelines. Figure 2.2 below shows the web interface of CancerPPD2 hosted at https://webs.iiitd.ac.in/raghava/cancerppd2.

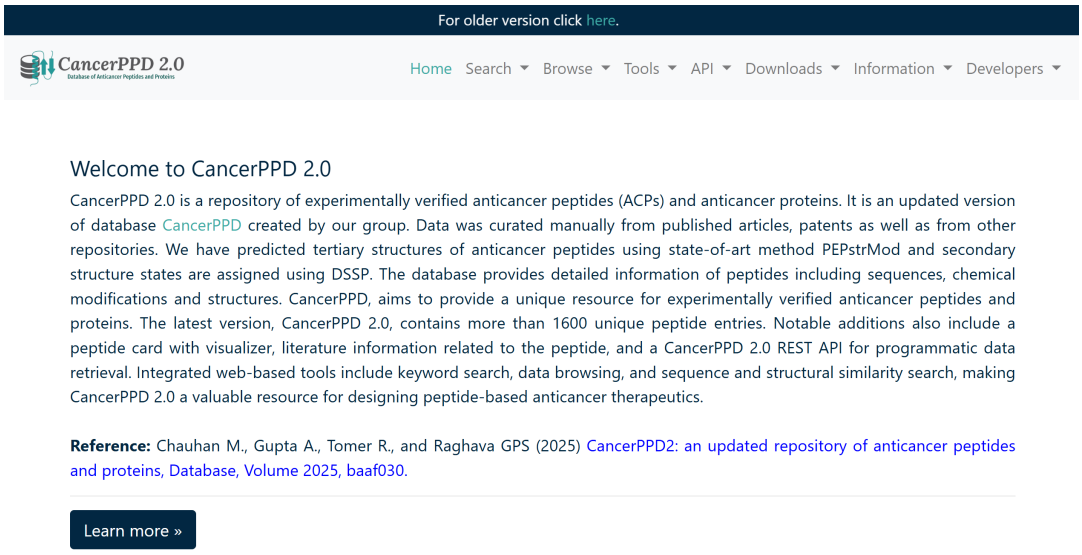


Fig. 2.2. *Web Interface of CancerPPD2*

2.4 Database Content

The database is primarily organized into two categories– Primary information and Secondary information. **Primary information** was manually curated from the literature directly and consists of various fields, including -

i. PMID
ii. Peptide sequence
iii. Name of the peptide
iv. Length of the peptide
v. Configuration (linear or cyclic)
vi. Chirality (L/D/Mix)
vii. chemical modification
viii. N-terminal modification
ix. C-terminal modification
x. Origin of the peptide
xi. Anticancer activity of peptide
xii. Tested cell lines
xiii. Assay types
xiv. Test time
xv. Cancer types
xvi. Target tissues

Figure 2.3 illustrates a sample entry from CancerPPD2.0, showcasing the primary details of a peptide. The entry includes the CancerPPD2 ID: `4189`, along with literature-curated annotations describing the peptide's characteristics.

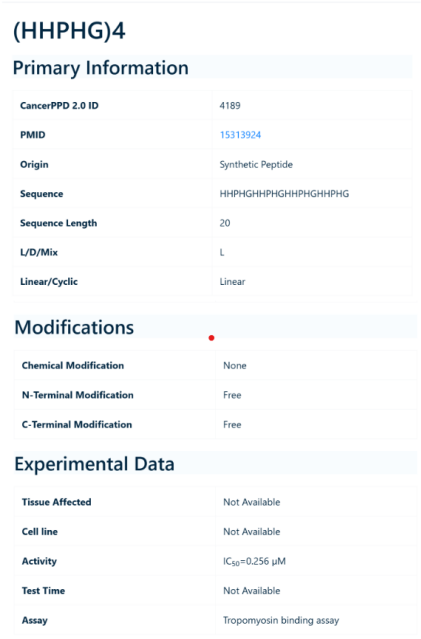


Fig. 2.3. *Example of a CancerPPD2.0 entry displaying primary information, including CancerPPD2 ID and literature-curated annotations.*

Secondary information is derived from the primary information and consists of structural details of the peptide like tertiary structures and simplified molecular-input line-entry system (SMILES). Figure 2.4 below shows the secondary information associated with CancerPPD2 ID `4189`, available on the database. We compiled all peptide structures and SMILES for an exhaustive database. Firstly, we mined the Protein Data Bank (PDB) to collect existing peptide/protein structures. Secondly, we used `PepstrMOD` [20], an updated version of PepStr [21] for predicting the structure of peptides. To the best of our knowledge, `PepstrMOD` is the only method that can predict the structure of peptides containing natural or chemically modified peptides. Thus, all peptides in our dataset were predicted using `PepstrMOD` whether they are natural peptides or chemically modified or N-/C-terminal-modified peptides. `PepstrMOD` accepts sequences in the range of 7 - 25 residues only. For peptides which contain more than 25 residues or less than 7 residues, we tweaked the lower and upper limit of `PepstrMOD` to accept peptides between 5 and 40 residues to include more natural and modified peptides. The structure of anticancer peptides/proteins obtained from UniProt was extracted from the cross-reference database '`AlphaFoldDB`'. For anticancer proteins whose structure is not available in '`AlphaFoldDB`' was predicted using software `I-TASSER` [22].

Fig. 2.4. *Secondary information associated with a CancerPPD2.0 entry - 4189, showing DSSP SMILES*

Each entry in CancerPPD2 is represented through a dedicated and dynamically generated '**Peptide Card**', offering users a comprehensive snapshot of all available data related to a specific anticancer peptide or protein. These cards consolidate key details including peptide origin, sequence, structural and functional annotations, and biological activity data, enabling researchers to access critical information at a glance. A prominent feature of the Peptide Card is the interactive 3D structural visualization, implemented using the NGL Viewer [23] (See Figure 2.5). This allows users to manipulate the three-dimensional conformation of peptides and proteins in real-time—zooming, rotating, and inspecting specific residues with high precision.

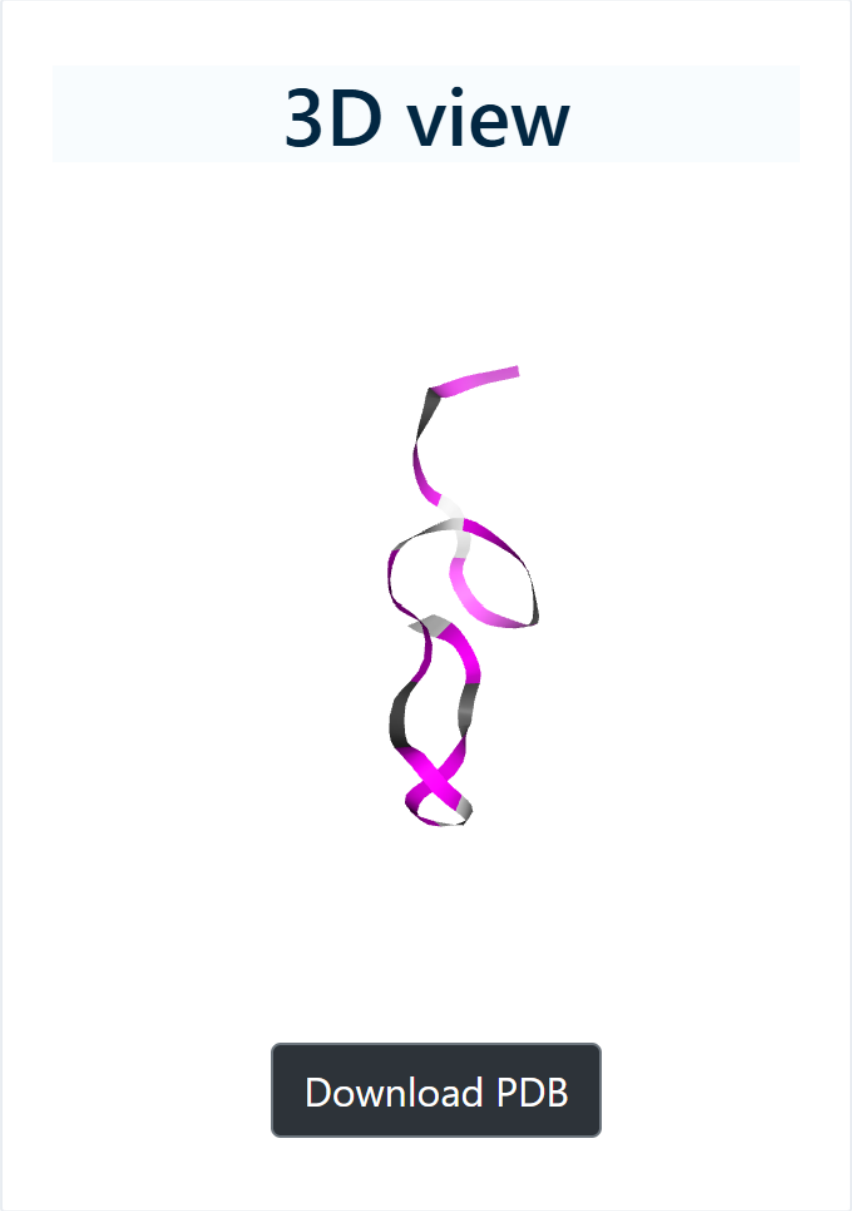


Fig. 2.5. *structural visualization of an anticancer peptide (ACP) for peptide entry 4189 using the NGL Viewer integrated within CancerPPD2.0.*

Additionally, each entry is enriched with curated literature references, including the title, journal source, and abstract retrieved on the go using the PubMed API, (see Figure 2.6) offering context and traceability to the original experimental evidence. Where available, cross-references are provided to external databases such as the Protein Data Bank (PDB), UniProt/Swiss-Prot, TrEMBL, and the Immune Epitope Database (IEDB), facilitating deeper exploration of structural, functional, and immunological data.

Literature

Title of the Paper:

Peptides derived from the histidine-proline domain of the histidine-proline-rich glycoprotein bind to tropomyosin and have antiangiogenic and antitumor activities.

DOI: 10.1158/0008-5472.CAN-04-0440

Authors:

Fernando Doñate, Jose C Juarez, Xiaojun Guan, Natalya V Shipulina, Marian L Plunkett, Ziva Tel-Tsur, David E Shaw, William T Morgan, Andrew P Mazar

Year and Journal:

2004 , Cancer research

Abstract:

The antiangiogenic activity of the multidomain plasma protein histidine-proline-rich glycoprotein (HPRG) is localized to its histidine-proline-rich (H/P) domain and has recently been shown to be mediated, at least partially, through binding to cell-surface tropomyosin in fibroblast growth factor-2-activated endothelial cells (X. Guan et al., Thromb Haemost, in press). HPRG and its H/P domain, but not the other domains of HPRG, bind specifically and with high affinity to tropomyosin. In this study, we characterize the interaction of the H/P domain with tropomyosin and delineate the region within the H/P domain responsible for that interaction. The H/P domain of HPRG consists mostly of repetitions of the consensus sequence [H/P][H/P]PHG. Applying an in vitro tropomyosin binding assay, we demonstrate that the synthetic peptide HHPHG binds to tropomyosin in vitro and inhibits angiogenesis and tumor growth in vivo. The affinity for tropomyosin increases exponentially upon multimerization of the HHPHG sequence, with a concurrent increase in antiangiogenic activity. Specifically, the tetramer (HHPHG)4 has significant antiangiogenic activity in the Matrigel plug model (IC50 approximately 600 nm) and antitumor effects in two syngeneic mouse tumor models. Thus, we show that a 16-mer peptide analogue mimics the antiangiogenic activity of intact HPRG and is also able to inhibit tumor growth, suggesting that cell surface tropomyosin may represent a novel antiangiogenic target for the treatment of cancer.

Fig. 2.6. *Literature references retrieved via the PubMed API, providing source links and publication details for ACP entries in CancerPPD2.0 for peptide entry 4189.*

Collectively, the Peptide Card interface in CancerPPD2 serves as a centralized and interactive information hub, streamlining access to multi-faceted data for both computational and experimental researchers in the field of anticancer therapeutics. Figure 2.7 below shows complete Peptide Card for peptide entry **1859** at CancerPPD2.

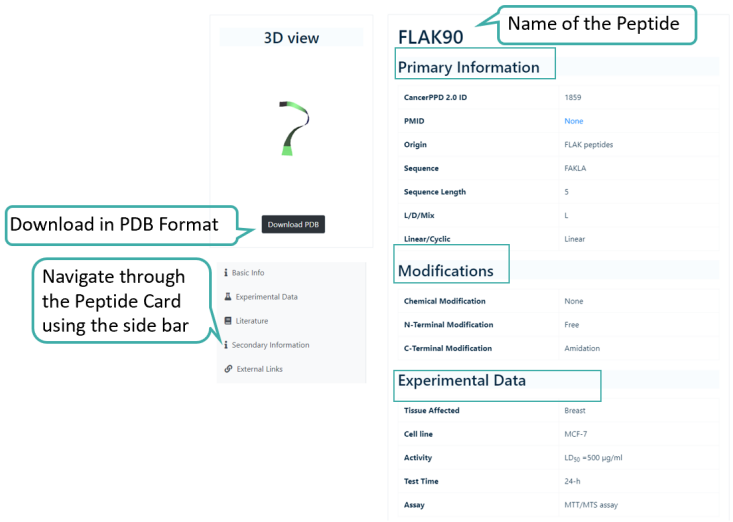


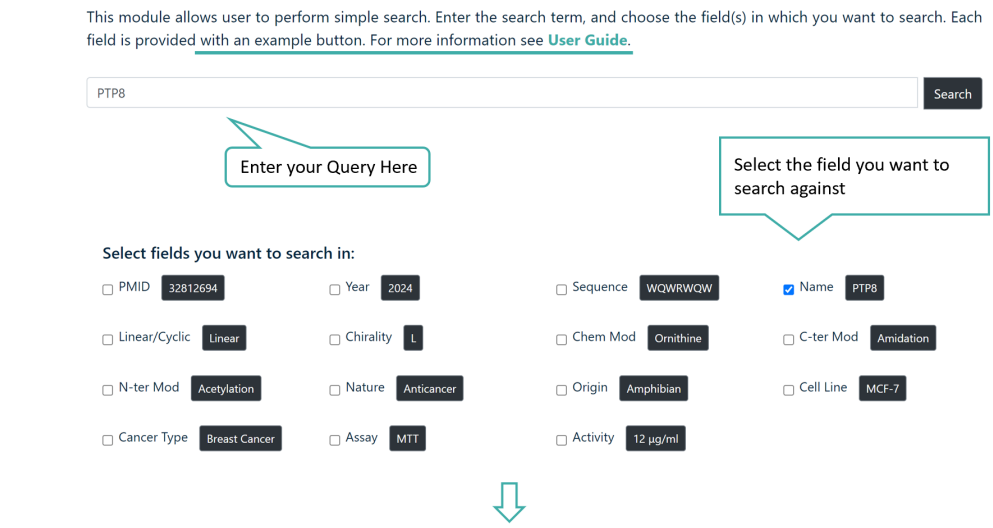
Fig. 2.7. *Representative "Peptide Card" (ID 1859) in CancerPPD2.0, integrating structural, functional, and bibliographic data in a single view.*

2.5 Data Retrieval

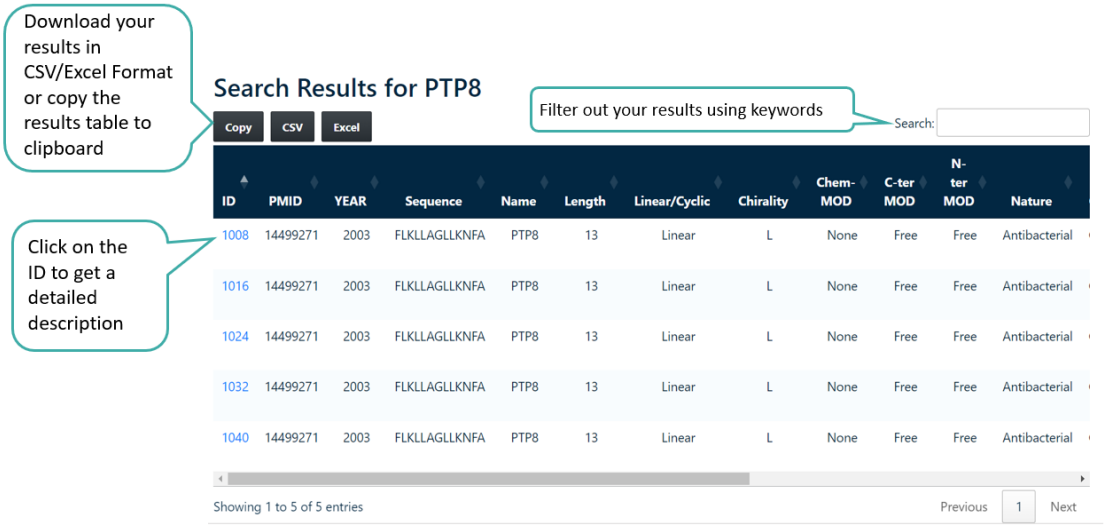
Different modules have been integrated into the web interface to facilitate data retrieval in an efficient way. These modules allow the users to perform targeted searches across the database through a streamlined and user-friendly interface. The database provides multiple modules for data retrieval including interactive search tools, RESTful APIs for programmatic access, and a dedicated downloads section for bulk data retrieval. The detailed description of each of the search modules is given below -

2.5.1 Basic Search

This module enables users to perform straightforward keyword-based searches across the database. Users can input a search term and specify one or more data fields (such as *Peptide ID*, *Sequence*, *Cancer Type*, or *Origin*) in which the term should be queried. Figure 2.8 below shows the basic search module and the retrieved search results.



(a) *Basic Search Module*

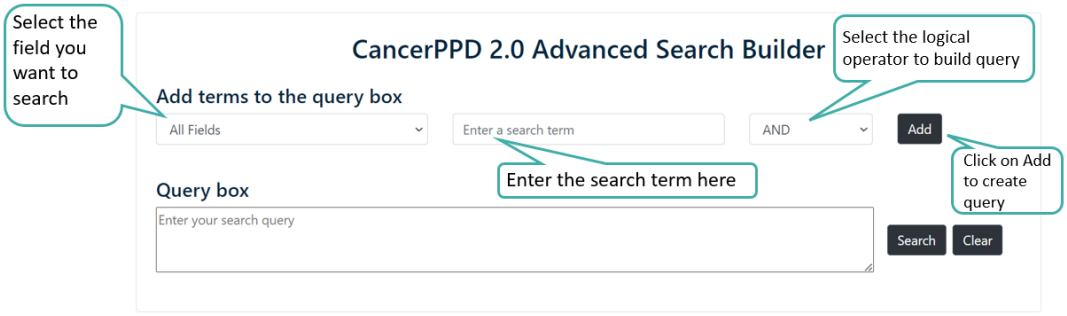


(b) *Search Results of Basic Search*

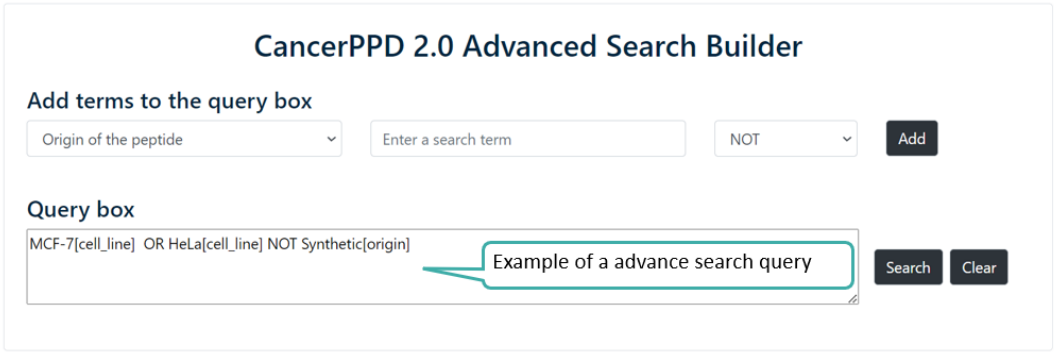
Fig. 2.8. *(a) and (b): Interface of the Basic Search Module in CancerPPD2.0*

2.5.2 Advanced Search

The Advanced Search module allows the user to construct complex, multi-parameter queries. It allows selection of multiple fields and supports the use of standard Boolean operators such as **AND** , **OR** and **NOT** to perform conditional searches. Users can iteratively build their queries using the "Add" button, allowing them to refine or customize their searches, and facilitating data retrieval across various fields, e.g., Sequence, Cancer Type, Cell Type etc. Figure 2.9 below shows the Advanced Search module and retrieved search results.



(a) *Advanced Search Module*



(b) *Search Results of Advanced Search*

Fig. 2.9. *(a) and (b): Interface of the Advanced Search Module in CancerPPD2.0*

2.5.3 Peptide Search

The **Peptide Search** module enables users to query the database using peptide sequence information. It supports two modes of sequence-based retrieval:

1. **Identical Sequence Search:** This option retrieves entries that match the input peptide sequence exactly, allowing users to identify peptides that are already cataloged in the database with identical sequences.
2. **Subsequence Search:** This option allows users to search for peptides that contain the input query as a contiguous subsequence, enabling the identification of peptides that share conserved sequence motifs or structural fragments.

These functionalities are particularly useful for researchers aiming to explore sequence redundancy, motif conservation, or functional analogs within the curated anticancer peptide dataset. Figure 2.10 below shows the Peptide Search module.

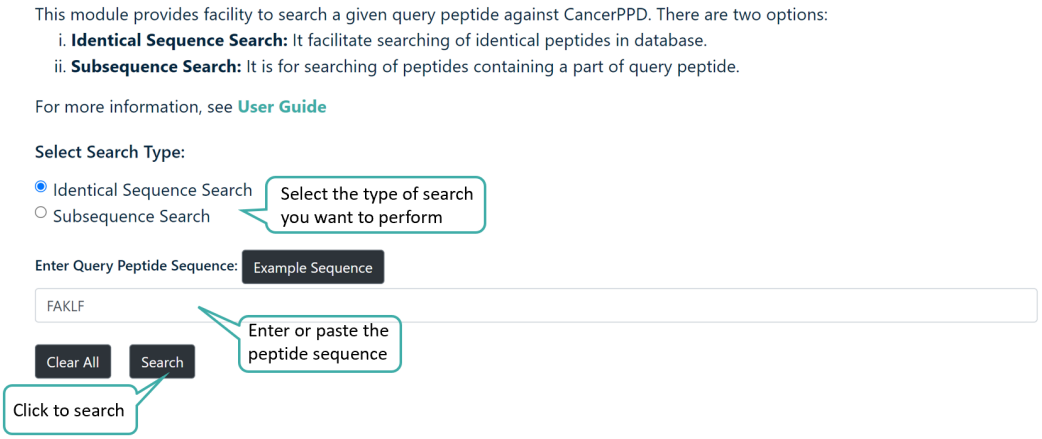


Fig. 2.10. *Peptide Search Module*

2.5.4 SMILES Search

While traditional sequence-based search modules facilitate exploration at the amino acid level, understanding the structural determinants of anticancer peptides (ACPs) often requires analysis at the chemical structure level. To support such investigations, CancerPPD 2.0 maintains the chemical structures of peptides in SMILES (Simplified Molecular Input Line Entry System) format, enabling atom- and bond-level analysis.

The SMILES Search module allows users to query the database using a SMILES string of a chemical structure. This feature supports four types of structure-based searches:

1. **Substructure Search:** Identifies database entries that contain the query SMILES as a substructure within the complete peptide molecule. This is useful for detecting conserved chemical moieties or pharmacophores.
2. **Exact Search:** Retrieves peptides whose SMILES representations are an exact match to the query. This ensures precise identification of chemically identical structures.
3. **Exact Fragment Search:** Searches for an exact match to the specified fragment within larger peptide structures, offering specificity in fragment-level analysis.
4. **Superstructure Search:** Returns entries where the query SMILES is a superstructure of peptides in the database, allowing the identification of peptides that are subcomponents of the query molecule.

Figure 2.11 below shows the Peptide Search module.

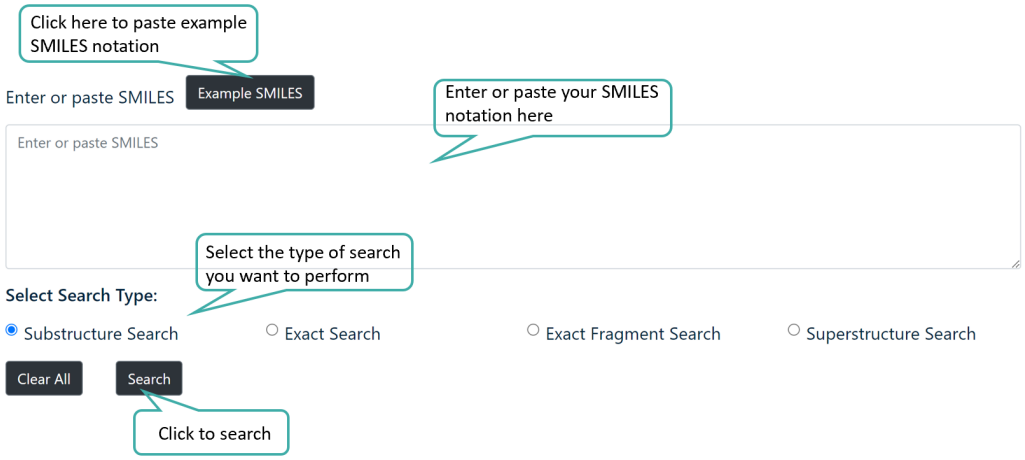


Fig. 2.11. *SMILES Search Module*

2.5.5 REST API

To facilitate programmatic data retrieval, a RESTful API has also been integrated into the database. Users can access data using simple REST URLs. The REST API returns responses in `JSON` (JavaScript Object Notation) format, which can be easily parsed and customized to meet specific data processing needs. Upon sending a request to the server the following HTTP response headers are returned by the CancerPPD 2.0 REST API:

Code	Description
200	The request was processed successfully.
400	Bad Request. Invalid data type.
404	Not Found. The requested data doesn't exist.
500	Internal server error. Most likely a temporary problem, but if the problem persists please contact us.

Table 2.1. *API Response Codes and Their Descriptions*

1. Query Fields

The CancerPPD 2.0 REST API provides access to its data through three distinct query fields: **cancer type**, **cell line**, and **peptide sequence** (See Table 2.2). Each field offers a diverse range of options to enable targeted and flexible data retrieval.Specifically, the *cancer type* query supports selection among 15 different cancer types. The *cell line* query field allows access to experimental data from 45 unique cell lines, chosen based on their relevance and representation within the database. Additionally, the *peptide sequence* field supports

refinement based on the nature of the peptide, offering two parameters: **Natural** and **Modified**, enabling users to distinguish between unaltered peptides and those that have undergone chemical modifications.

Query Field (dataType)	**Parameter (dataValue)**	**Description**
Cancer Type (cancer_type)	Breast Cancer, Lung Cancer, Ovarian Cancer, Cervical Cancer, Colon Cancer, Skin Cancer, Blood Cancer, Prostate Cancer, Liver Cancer, etc.	Users can access data corresponding to a particular cancer type. This will return all the entries for the selected cancer type.
Cell Line (cell_line)	MCF-7, A-549, HeLa, PC-3, HepG-2, MDA-MB-231, HCT-116, Jurkat, K-562, MDA-MB-435S, HT-1080, DU-145, HL-60, Caco-2, HMLER, etc.	Users can access data corresponding to a particular cell line. This will return all the entries for the selected cell line.
Peptide Sequence (seq)	Natural, Modified	Users can access data based on their selected parameter. Selecting "Natural" retrieves entries with sequences composed solely of natural amino acid residues without modifications, while "Modified" returns sequences with non-natural residues or chemical modifications.

Table 2.2. *Query fields, parameters, and descriptions supported by the CancerPPD 2.0 REST API*

2. Return Fields

Once the request has been processed successfully, the CancerPPD 2.0 REST API returns the data in `JSON` format. The response consists of the following 16 fields:

Return Field	Description
CancerPPD 2.0 ID	Unique identifier in the CancerPPD 2.0 database for that entry.
PMID	Article PMID corresponding to that entry.
Year	Year of publication of that entry.
Sequence	Sequence of the peptide.
Name	Name of the peptide.
Length	Length of the peptide.
Linear/Cyclic	Conformation of the peptide.
Chirality	Stereochemistry of the peptide.
Chemical Modifications	Whether the peptide contains any non-natural residues or any other chemical modifications.
C-Ter Modifications	Whether the C-terminal end of the peptide contains an entity or it is free.
N-Ter Modifications	Whether the N-terminal end of the peptide contains an entity or it is free.
Cell Line	Anticancer property of the peptide tested against which cell line.
Cancer Type	The cancer type corresponding to that cell line.
Assay	Type of assay used to measure the anticancer activity of the peptide.
Test Time	Time after which the anticancer activity was measured or found to be significant.
Tissue Affected	The tissue affected corresponding to the cancer type.

Table 2.3. *Description of return fields in the JSON response from CancerPPD 2.0 REST API*

3. cURL

Data can be accessed programmatically using command-line tools such as **cURL**. cURL (Client URL) is a versatile command-line tool and library used for transferring data with URLs across various protocols including HTTP and HTTPS. It allows users to send requests and receive responses directly from the terminal or within scripts, making it ideal for programmatic access to web APIs.

For example, to fetch data on the cell line `A-549` from CancerPPD 2.0, the following cURL command can be used:

```
curl -X GET "https://webs.iiitd.edu.in/cancerppd2/api/api.
    ↪ php?dataType=cell_line&dataValue=A-549"
```

This command initiates a `GET` request to the REST API endpoint, retrieving the requested data in `JSON` format. Users can modify the query parameters (`dataType` and `dataValue`) to access different datasets according to their re-

search needs. The integration of `cURL` facilitates easy testing of API endpoints and can be embedded in automated workflows for high-throughput data analysis.

4. wGET

`wget` is another widely used command-line utility designed for downloading files from the web, supporting protocols such as `HTTP`, `HTTPS`, and `FTP`. It is particularly prevalent in Linux and Unix-based environments and is useful for automating data retrieval in batch processes or shell scripts. For instance, to download data related to the cell line `A-549` from the CancerPPD 2.0 REST API, the following command can be used:

```
wget "https://webs.iiitd.edu.in/raghava/cancerppd2/api/api
   ↪ .php?dataType=cell_line&dataValue=A-549"
```

Usage Instructions

To retrieve data programmatically via the CancerPPD 2.0 REST API, users select the desired query field along with its corresponding search term and initiate the request by clicking the **Execute** button. This action generates the API call and returns the data matching the specified criteria. Upon clicking **Execute**, the system displays the corresponding API request details, including the `cURL` command, `wget` command, request URL, and the server response. Figure 2.12 below shows the usage of the CancerPPD 2.0 REST API interface, illustrating how users can construct queries, execute requests, and view results such as cURL/wget commands and server responses. To download the data directly in your `python` environment use the following code.

```
import requests
import pandas as pd
#Define the URL for the API request
url = "https://webs.iiitd.edu.in/raghava/cancerppd2/api/
   ↪ api.php?dataType=cancer_type&dataValue='Fibrosarcoma
   ↪ '"
#Send a GET request to the API
response = requests.get(url)
#Check if the request was successful (status code 200)

if response.status_code == 200:
    #Print the JSON response from the API
    print(response.json())

    #Parse the JSON data and store it in a pandas
       ↪ DataFrame
    json_data = response.json()
    df = pd.DataFrame(json_data['data'])
```

```
    #Visualize the DataFrame
    print(df.head()) #Display the first few rows of the
        ↪ DataFrame

else:
    #Print an error message if the request was not
        ↪ successful
    print('Error:', response.status_code)
```

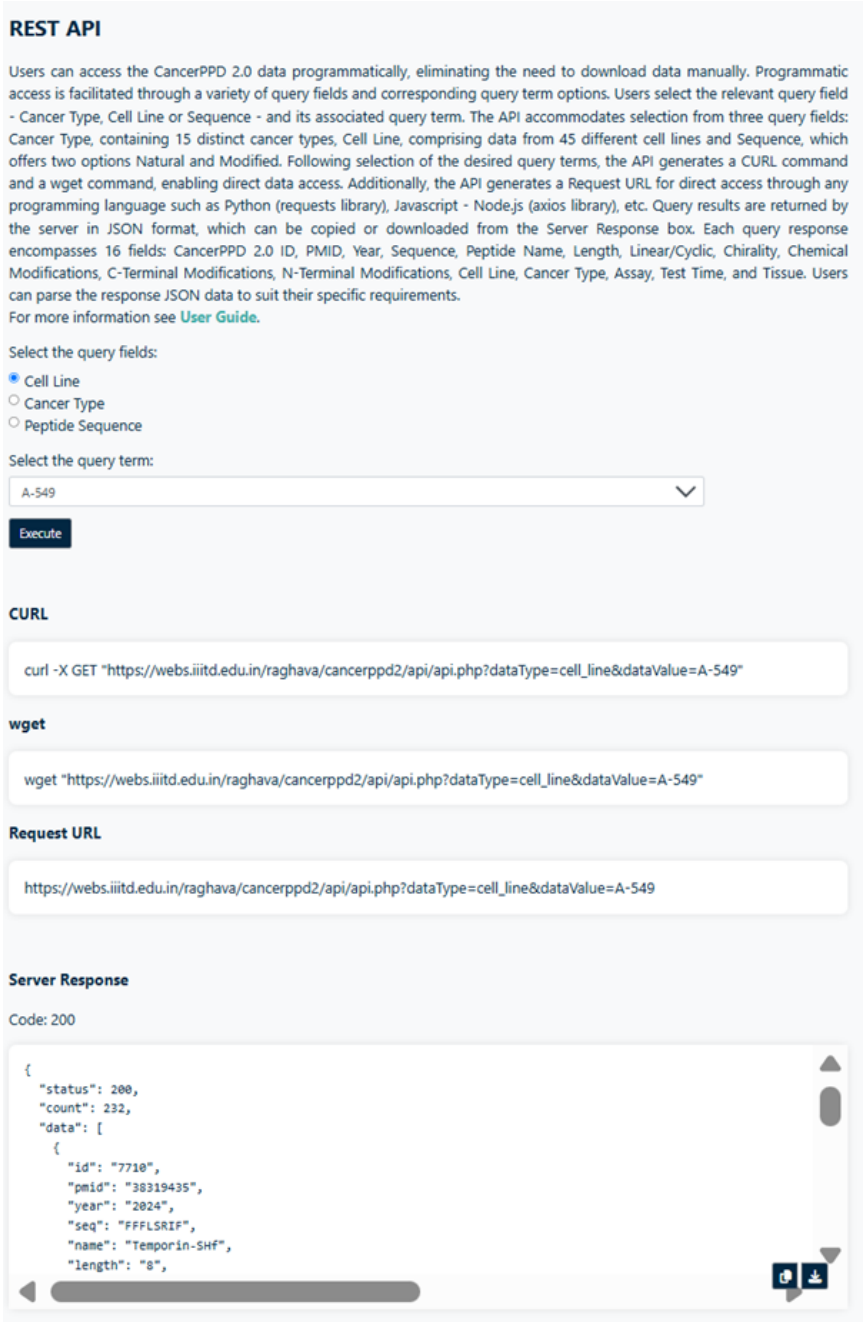
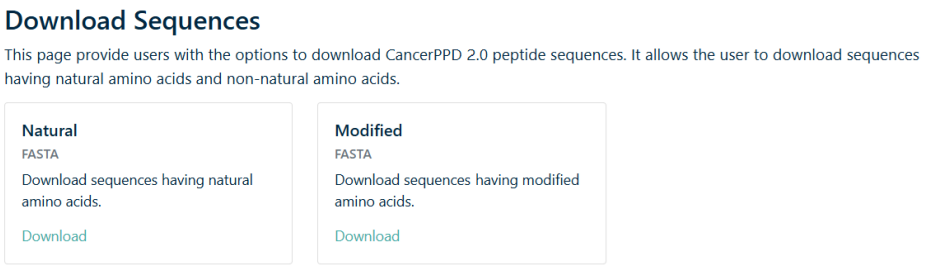


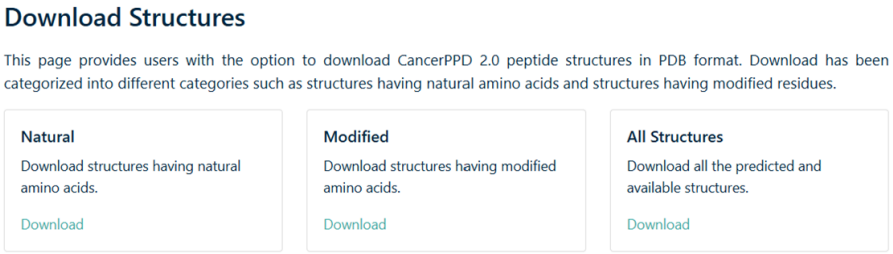
Fig. 2.12. *Usage of REST API on CancerPPD2 web interface*

2.6 Downloads

The **Downloads** section of CancerPPD 2.0 provides users with direct access to key data resources in standardized formats. Users can download natural and modified peptide sequences separately in `FASTA` format (See Figure 2.13a). Additionally, 3D structural data of peptides and proteins are available in `PDB` format to facilitate structural and computational analyses (See Figure 2.13b). For literature curation, we have included `PDF` files of publicly accessible research articles from PubMed that were used to annotate anticancer peptides in the database (See Figure 2.13c). Users are also provided the option to download or view the results of their search queries in `CSV` or `Excel` formats. These resources are intended to support both experimental and computational investigations in peptide-based cancer research.

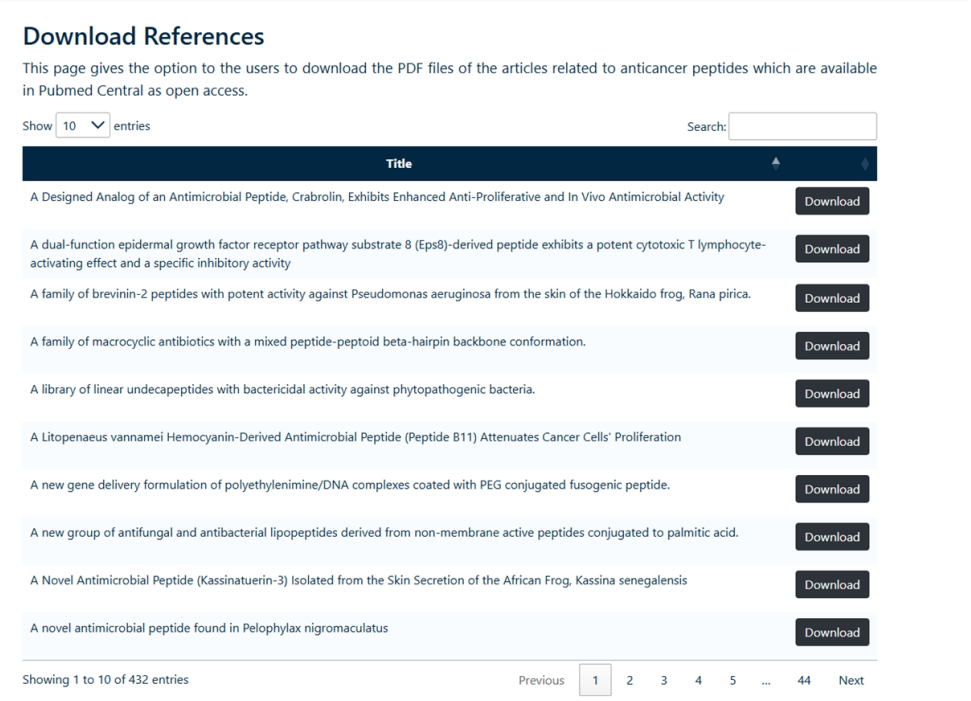


(a) *Peptide sequence downloads*



(b) *Predicted structure downloads*

Fig. 2.13. *Interface of the Download section in CancerPPD2.0: (a) Peptide sequence downloads, (b) predicted structure downloads, and (c) literature reference downloads* *(continued on next page).*



(c) *Literature reference downloads interface in CancerPPD2.0.*

2.7 Browsing Tools

We have developed an intuitive browsing interface to enable the structured retrieval of information from CancerPPD2. This user-friendly module allows users to explore anticancer peptides (ACPs) and proteins based on key fields, including *peptide/protein-wise*, *tissue type*, *cancer cell line type*, *year of discovery*, *assay type*, and *peptide length*. ACPs can also be browsed according to their *chemical modifications*, such as *N-terminal* and *C-terminal* modifications. In addition to peptide and protein-level information, users can access associated *clinical trial data*, with cross-links to relevant *NCT studies*. A dedicated section lists anticancer proteins found in *UniProt*, with direct links to their corresponding UniProt entries. The database encompasses **392** cancer cell lines derived from **28** distinct tissue types, enabling users to browse peptides tested against specific cancer models. All retrieved data are presented in an interactive tabular format, allowing for easy navigation and interpretation. Figure 2.14 below illustrates the browsing of all anticancer peptides.

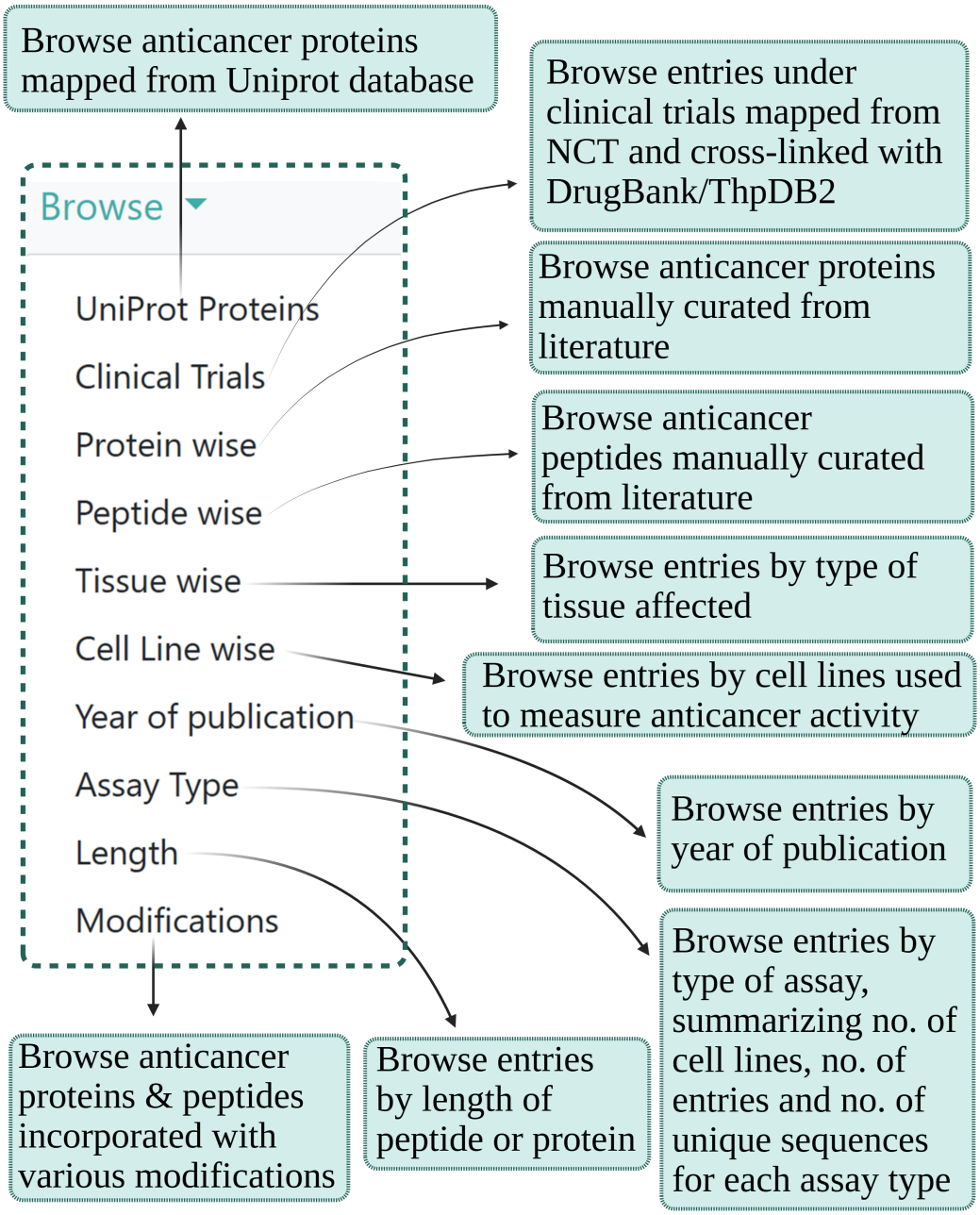


Fig. 2.14. *Browse section in CancerPPD2*

2.8 Analysis Tools

CancerPPD 2.0 incorporates a suite of web-based analytical tools to facilitate comprehensive peptide characterization. For sequence similarity analysis, the database integrates the **Basic Local Alignment Search Tool** (BLAST) [24], enabling users to identify anticancer peptides (ACPs) with high sequence homology to a given query. Users can run a BLAST query against the CancerPPD 2.0 database. After submission of job it returns the list of peptides similar to the query peptide. The server also provides options to choose different parame-

ters like weight matrix and expectation value. Figure 2.15 below shows BLAST interface for sequence similarity in CancerPPD2 database.

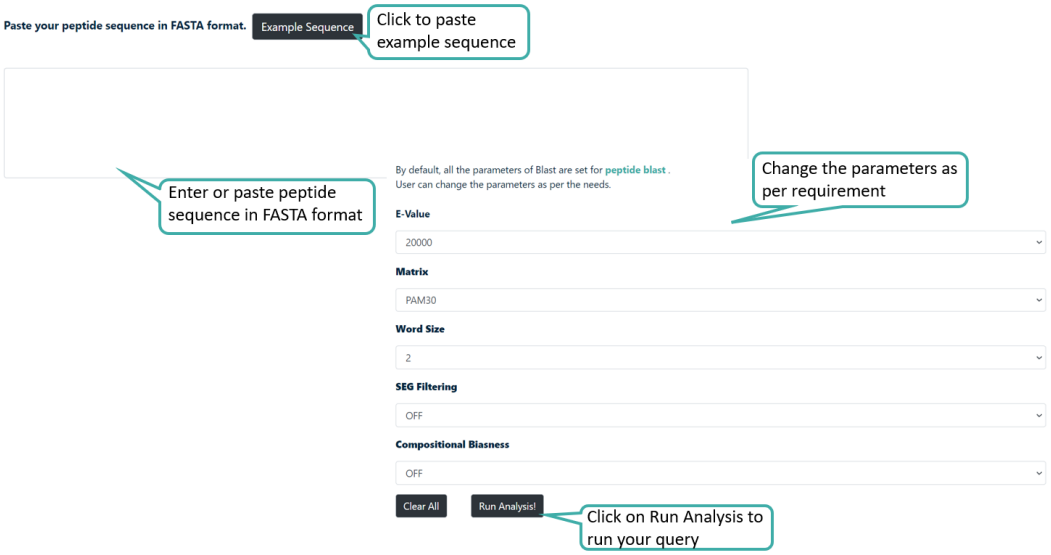


Fig. 2.15. *BLAST Module*

In addition, a **Smith–Waterman** algorithm-based tool [25] has been implemented for more sensitive local alignment. Users can run a Smith-Waterman search query against the CancerPPD 2.0 database (see Figure 2.16). After submission of job it returns the list of peptides.

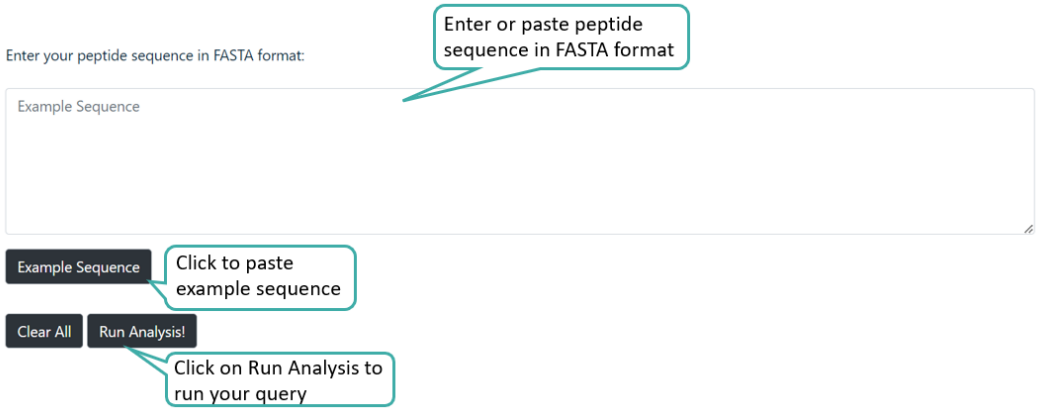


Fig. 2.16. *Smith-Waterman Module in CancerPPD2.0*

To assist in functional mapping, a **Peptide-mapping** module allows users to scan their query proteins for known ACPs present in CancerPPD. This enables users to locate regions within proteins that correspond to experimentally validated anticancer peptides. Users can select either SuperSearch to search for query PROTEIN sequence against peptides of CancerPPD or SubSearch to search for query PEPTIDE sequence against the peptides of CancerPPD 2.0. Figure 2.17 shows the Peptide mapping module in CancerPPD2.

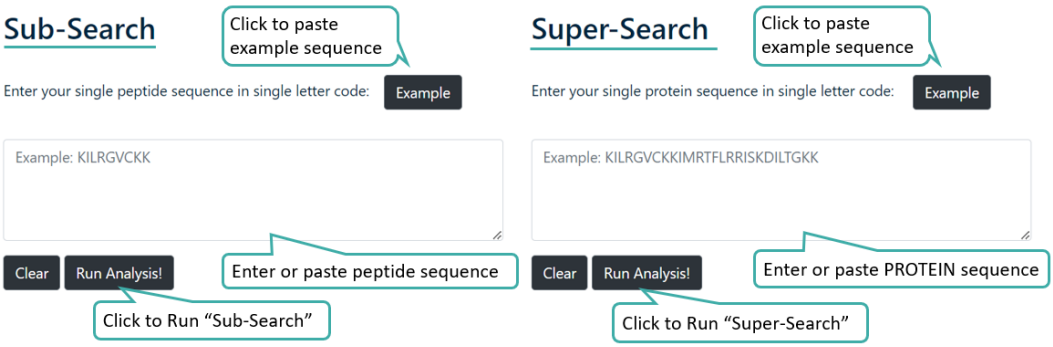
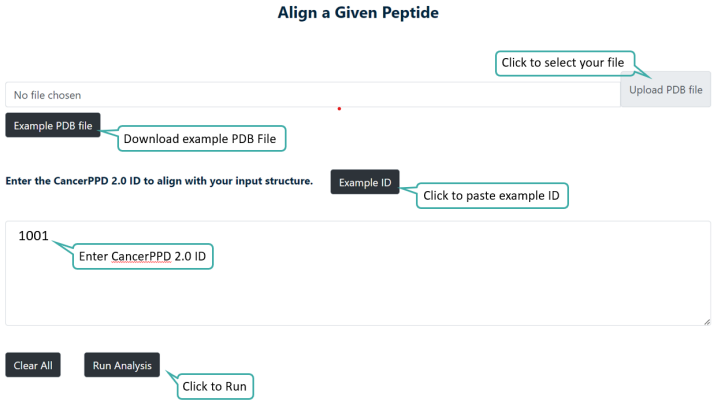
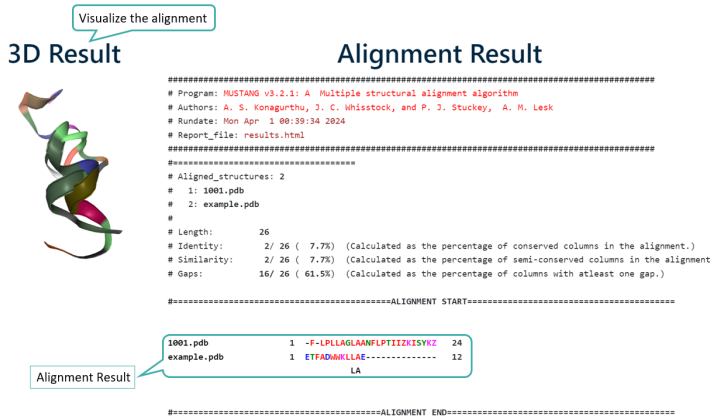


Fig. 2.17. *Peptide Mapping Module*

For **structural alignment**, the platform integrates MUSTANG [26], which performs structure alignment of two peptide or protein structures. This functionality provides insights into the structural conservation among anticancer peptides. User can align their `PDB` structure with any of the CancerPPD 2.0 Structures (see Figure 2.18).



(a) *Structure Alignment Module: Interface for aligning 3D structures of anticancer peptides.*



(b) *Result of Structure Alignment: Visualization of structural superposition between anticancer peptides.*

Fig. 2.18. *(a) and (b): Interface and output of the Structure Alignment Module in CancerPPD2.0.*

2.9 Results

The updated version of our database, CancerPPD 2.0, comprises a total of **6,521** entries, representing a substantial expansion over the previous version. The entries have been curated from scientific literature and publicly available databases, and include every available experimental and annotation detail retrievable at the time of compilation. The database now includes **5,919** anticancer peptides (ACPs), **541** anticancer proteins, and **61** entries classified as 'Unknown' due to missing amino acid sequence and/or length information. Despite lacking sequence data, the *'Unknown'* category retains critical metadata such as reported anticancer activity, cell lines used for testing, assay types, and associated tissue origins [27–29].

Recognizing the diversity in biological activity of ACPs across different experimental contexts, CancerPPD 2.0 records multiple entries for the same ACP when evaluated under distinct conditions, including variations in cell lines, IC_{50} values, assay protocols, or tissue sources. This decision reflects the complex and

context-specific behavior of ACPs and enhances the granularity of the dataset for downstream analyses.

To support translational relevance, the database integrates **47** entries linked to clinical trials, each mapped to unique NCT identifiers, providing users with direct connections to trial metadata. Furthermore, we document experimental results on **392** unique cancer cell lines derived from **28** distinct tissue types, offering a comprehensive landscape of cellular models employed in ACP evaluation.

Importantly, in response to the therapeutic challenges associated with peptide stability, we have expanded the dataset to include **781** chemically modified peptides, reflecting ongoing efforts to improve bioavailability and therapeutic potential of ACPs.

In CancerPPD 2.0, we have incorporated **2,661** entries corresponding to **1,005** unique natural anticancer peptides and proteins, enriched with structural information wherever available. Among these, experimental three-dimensional (3D) structures for **111 ACPs** were retrieved directly from the Protein Data Bank (PDB). For the remaining peptides lacking experimentally determined structures, we employed state-of-the-art structure prediction tools including `PepstrMOD`, `AlphaFold`, and `I-TASSER`, predicting structures for **541**, **401**, and **52** ACPs, respectively. It is important to note that structural modeling could not be performed for peptides shorter than five amino acids, due to limitations in conformational prediction reliability for very short sequences.

Terminal and chemical modifications play a pivotal role in enhancing the pharmacological properties of anticancer peptides (ACPs). In CancerPPD 2.0, special emphasis has been placed on capturing and categorizing these modifications due to their profound impact on peptide stability, bioavailability, protease resistance, cell penetration, and receptor binding affinity.

N- and C-terminal modifications—such as *acetylation, amidation, and pegylation*—are among the most common strategies employed to improve peptide half-life in biological systems. These modifications shield the peptide termini from exopeptidase degradation, thereby enhancing metabolic stability. The database includes **3,018** entries covering **742** unique ACPs with such terminal modifications. These modified peptides often demonstrate superior pharmacokinetic profiles, making them attractive candidates for therapeutic development. We were able to predict the 3D structures for about **673** of these ACPs using `PepstrMOD`, facilitating further exploration of structure–activity relationships.

In addition to terminal modifications, chemical modifications involving side chains or non-natural residues offer opportunities to fine-tune peptide function and expand their chemical diversity. CancerPPD 2.0 catalogs **781** entries for ACPs bearing non-terminal chemical modifications, which include substitutions or incorporations of non-canonical amino acids such as *1-amino-isobutyric acid, -naphthylalanine, norleucine*, and *ornithine*. These residues often confer enhanced helicity, membrane interaction, or target specificity. However, due to the structural complexity and absence of universal force field parameters, re-

liable 3D structure prediction was achieved for only **33** such peptides using `PepstrMOD`. Along with the complex modifications, some of the structures were also not predicted due to lack of sequences, as their complex structures were given as figures in the source publications [30–33].

Many anticancer peptides entering preclinical or clinical pipelines incorporate such modifications to improve efficacy while minimizing immunogenicity and off-target effects. By systematically annotating and structurally characterizing these modifications, CancerPPD 2.0 provides a critical resource for rational peptide engineering and drug discovery.

Figure 2.19 provides a detailed visualization of the CancerPPD 2.0 database, showcasing the diversity and distribution of anticancer peptides. The **top left bar chart** illustrates the number of peptide entries across different length bins. The **adjacent donut chart** displays the distribution of entries among the 10 most common cancer types. The **lower left bar chart** presents the number of entries for the 10 most frequently studied cancer cell lines. The **adjacent bar chart** shows the number of peptide entries across various year bins. The **given pie chart** shows the anticancer activity of peptides/proteins with cancer type.

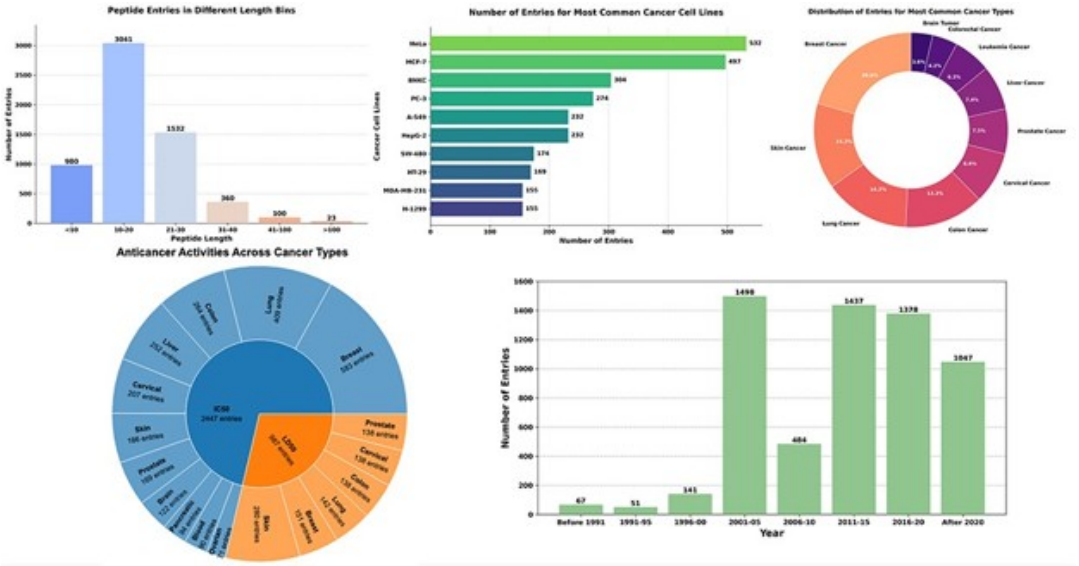


Fig. 2.19. *Overview of CancerPPD 2.0: Distribution of peptide lengths, cancer types, cell lines, publication years, and activity annotations.*

2.10 Comparison with previous version

CancerPPD2.0 represents a comprehensive upgrade over the original **CancerPPD** database, which was first released in **2014** as a curated repository of peptides and proteins exhibiting anticancer activity. Since its initial development, the field of *anticancer peptide (ACP)* research has witnessed considerable growth, necessitating a substantial update to reflect new findings and therapeutic advances.

In this revised version, we have significantly expanded the database by in-

corporating **2,612** additional entries, curated from recent peer-reviewed publications and patent literature (sourced from **PubMed** and **Patent Lens**). Each new entry is enriched with *primary and secondary information*, including **sequence details**, **activity data**, **assay conditions**, **structural models**, and relevant **cell line information**.

One of the major enhancements involves a marked increase in chemically modified entries—CancerPPD2.0 now contains **491** such entries, compared to only **290** in the previous version. This reflects the increasing emphasis on *peptide modification strategies* for improving therapeutic potential. Furthermore, the current version introduces **47** entries linked to clinical trials, each annotated with corresponding NCT identifiers, thereby bridging the gap between preclinical data and translational research.

To improve cross-platform utility and integration with other biomedical resources, we have also included cross-references to UniProt and NCT, enhancing the database's interoperability and data traceability. Additionally, a new dataset of *anticancer proteins and peptides* retrieved directly from UniProt under the *'anticancer'* annotation has been incorporated, further extending the coverage of the database.

One of the major new features introduced in CancerPPD2.0 is the implementation of a RESTful Application Programming Interface (API), which was absent in the original version of the database. This API has been developed to facilitate *programmatic access* to the database content, enabling seamless integration with computational pipelines, bioinformatics tools, and third-party applications.

A concise overview of the differences between the previous and updated versions is provided in Table 2.4, highlighting key advances in both *data volume* and *content diversity*. Collectively, these updates position CancerPPD2.0 as a significantly more comprehensive and valuable resource for the anticancer peptide research community.

Keyword	**CancerPPD**	**CancerPPD 2.0**	**Total Entries**
Anticancer peptides	3438	2481	5919
Anticancer proteins	121	420	541
Cell lines	249	226	392
Assays	16	41	49
Modified peptide entries	290	491	781
L-peptides	3274	2373	5647
D-peptides	26	31	57
Mixed peptides	178	183	361
Tissue types	21	24	28
Unique peptide entries	600	939	1539

Table 2.4. *Comparison of entries in CancerPPD and CancerPPD 2.0 across various key categories.*

With a total of **6,521** curated entries, CancerPPD2.0 offers enhanced coverage of anticancer peptides and proteins, incorporating both recent discoveries and detailed experimental metadata. The inclusion of newly identified ACPs from the latest literature ensures that the database remains current and comprehensive. Researchers can access the data through user-friendly web downloads or via the newly implemented API for programmatic access, enabling greater flexibility in usage. We anticipate that this updated version will serve as a valuable and versatile resource for researchers engaged in peptide-based anticancer drug discovery and development.

2.11 Discussion

Anticancer peptides (ACPs) are characteristically *cationic*, *amphipathic*, and *hydrophobic*, properties that are critical for their biological activity and therapeutic potential [7]. These peptides exhibit efficient tissue penetration and rapid cellular internalization, attributes that enable them to reach intracellular targets and exert potent cytotoxic effects [4, 7]. A key factor contributing to their selectivity is the anionic nature of cancer cell membranes, which arises from elevated levels of *phosphatidylserine*, *sialylated gangliosides*, and *heparan sulfate proteoglycans*. This negative surface charge facilitates the preferential binding of positively charged ACPs to cancer cells over normal cells, leading to membrane disruption and induction of cell death via necrosis or apoptosis [34].

In addition to their direct cytotoxic effects, ACPs have been shown to possess immunogenic properties, which may contribute to the stimulation of antitumor immune responses, further enhancing their therapeutic utility. The growing interest in ACPs as next-generation therapeutics is reflected in the increasing number of peptide-based candidates entering preclinical evaluation and various phases of clinical trials [8].

In light of this rapid advancement and the expanding relevance of ACPs in cancer therapy, CancerPPD2.0 was comprehensively updated to serve as a centralized, curated resource for researchers. By integrating structural, biochemical, pharmacological, and clinical data, the database aims to facilitate the rational design, optimization, and translational application of anticancer peptides. We envision CancerPPD2 as a valuable platform that will support both fundamental research and drug discovery pipelines, ultimately contributing to efforts aimed at addressing the global cancer burden.

2.12 Limitations

In the latest update, we have significantly increased the number of entries compared to the previous version, while incorporating enhanced features and a more user-friendly interface. To improve accessibility, we introduced a RESTful API, enabling seamless programmatic access to the data alongside the existing web-based interface, thereby catering to both computational and experimental researchers.

Despite these advancements, certain limitations remain. One of the major constraints is our inability to reliably predict the tertiary structures of peptides with complex chemical modifications, such as methoxylation, biotinylation, and hydroxamic acid substitutions. This limitation arises primarily due to the lack of dedicated force field parameters in current modeling tools, which hampers accurate prediction of non-canonical residues or elaborate chemical modifications.

Prediction Limitations for Peptide and Protein Structures

To predict the structures of peptides and proteins that were not readily available in the Protein Data Bank (PDB), we employed tools such as `PepstrMOD` and `I-TASSER`. While these tools are effective for a broad range of sequences, certain specific structures could not be predicted due to:

- Presence of complex modifications
- Intrinsic tool limitations
- Unusual configurations or non-standard residues

Overall, **1,314 structures** are available in CancerPPD2, of which:

- **1,258** were predicted using `PepstrMOD` or `I-TASSER`
- **56** were retrieved directly from PDB

We were unable to predict structures for a total of **298 entries**. Below is a breakdown of the major causes:

Natural Peptides

- **Sequence Not Available:** Some entries lacked sufficient sequence data.
- **Length < 5 residues:** `PepstrMOD` requires a minimum of 5 residues and a maximum of 25 for structure prediction.

Proteins

- **Mixed Chirality:** `I-TASSER` was unable to predict structures with mixed chirality—an acknowledged limitation of the tool.

Chemical Modifications

While `PepstrMOD` supported many chemically modified structures, several unique modifications presented challenges:

- **Non-natural residues:**
  - 2-Aminoisobutyric Acid

  - Aminohexanoic Acid
  - Dansylglycine
  - Selenocysteine
  - Pyroglutamic Acid
  - β-3-benzothienyl-1-Ala
  - Cyclohexyl-1-carboxylic acid

- **Terminal Modifications:** While common modifications like *acetylation* and *amidation* were handled by `PepstrMOD`, others posed issues, such as:
  - Folic Acid (N-terminal)
  - 1-octanoyl (N-terminal)
  - Rhodamine B – GABA (N-terminal)
  - Biotinylated (N-terminal)
  - Methoxylation (C-terminal)
  - Fluoromethyl ketone (C-terminal)
  - Leucinol (C-terminal) with 1-octanoyl (N-terminal)

Table 2.5 shows summary of peptide entries in CancerPPD2 for which three-dimensional structural information could not be predicted.

Structure Type	**Total Count**
Natural Peptides	25
Proteins	3
Modified Peptide Structures	162
Terminal Modifications	108
Total	**298**

Table 2.5. *Tabular summary of the total unpredicted structures*

Additionally, several peptides described in the literature lack complete sequence information or are represented only as schematic diagrams, limiting their inclusion in structural prediction workflows. We aim to address these limitations in future versions of CancerPPD by incorporating more robust modelling tools, expanding coverage of complex modifications, and curating more mechanistic and quantitative data, as such information becomes available in the literature.

Chapter 3

Utility of the Database

The CancerPPD2 database serves as a comprehensive, centralized platform for anticancer peptides (ACPs) and proteins, with a particular emphasis on chemically modified peptides. It offers manually curated annotations from peer-reviewed literature and publicly available resources, ensuring data accuracy, reproducibility, and biological relevance. This updated version builds on the widely-used CancerPPD (2015) developed by Tyagi et al., which has since become a foundational resource in the peptide therapeutics community.

Over the last decade, there has been a rapid expansion in the development of computational tools and machine learning models for ACP prediction. Nearly all major ACP prediction methods—such as AntiCP [11], ACP-DRL [12], ACP-ML [14], ACPPfel [15], AntiCP2 [16], MLACP 2.0 [17] etc., have relied on the CancerPPD dataset to train and benchmark their models. This widespread usage underscores CancerPPD's critical role in cancer peptide bioinformatics, both as a **benchmark dataset** and as a **knowledge base for therapeutic discovery**.

However, many recently discovered and experimentally validated ACPs especially those containing chemical modifications, D-residues, terminal modifications, and -natural amino acids—remain scattered across literature without integration into any centralized platform. This fragmentation posed a significant barrier to researchers seeking comprehensive and modification-aware datasets.

To address this gap, CancerPPD2 was developed as an updated and expanded repository that not only integrates new ACPs but also incorporates annotations for chemical modifications at the sequence and structural level. The database thus offers several important applications:

- **Dataset Development for ML/AI Pipelines**: Researchers can utilize CancerPPD2 to build more representative training datasets for machine learning models capable of capturing modification-aware biological activity.
- **Design and Optimization of Chemically Modified Peptides**: The modification-specific annotations allow peptide engineers to rationally design analogs with enhanced therapeutic potential, such as improved serum stability, selectivity, or reduced toxicity.
- **Benchmarking and Evaluation**: The curated and balanced dataset serves as a benchmark resource for evaluating the performance of newly developed computational predictors of anticancer activity.
- **Integration with Cheminformatics Tools**: With the inclusion of SMILES representations and other structura represtations, CancerPPD2

supports cheminformatics-driven applications such as virtual screening, QSAR modeling, and molecular docking.

Overall, CancerPPD2 bridges the gap between rapidly accumulating experimental data and the need for structured, accessible, and computationally usable peptide resources. It is poised to serve as a catalyst for future research in anticancer peptide discovery and design, particularly in the emerging area of modification-aware therapeutics. Figure 3.1 illustrates the *timeline of CancerPPD database development* and its pivotal role in advancing peptide-based drug discovery. By providing a high-quality, manually curated, and biologically relevant benchmarking dataset, the database has significantly accelerated the development of numerous in silico prediction tools. These tools, in turn, have enhanced the rational design and evaluation of anticancer peptides, fostering faster and more reliable therapeutic discovery in the peptide research community.

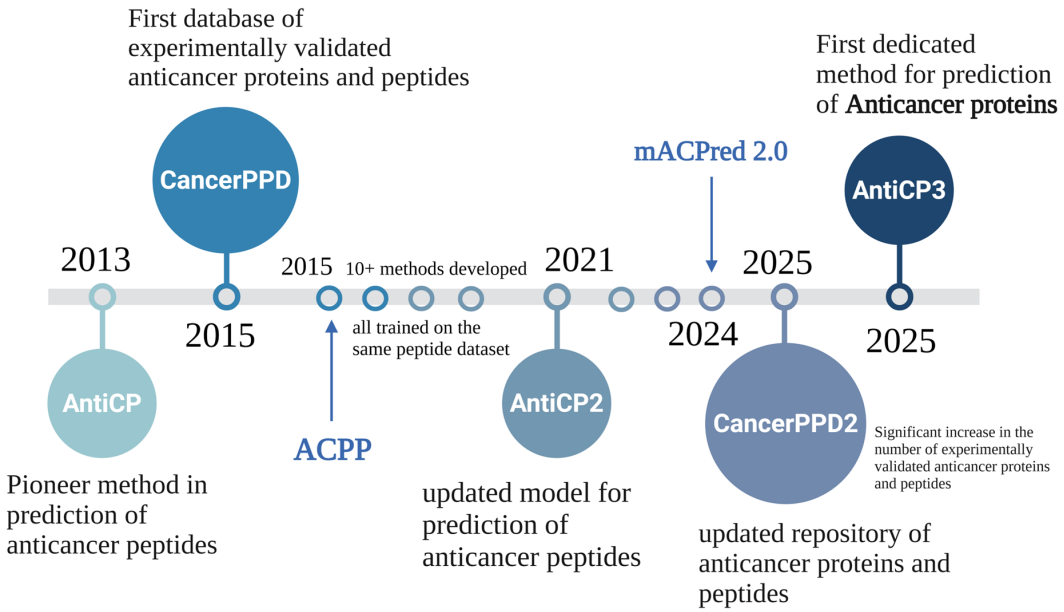


Fig. 3.1. *Timeline illustrating the development of in silico tools for peptide-based anticancer drug discovery.*

Chapter 4

Prediction of chemically-modified Anticancer peptides

4.1 Introduction

Although anticancer peptides (ACPs) hold tremendous potential as candidates for cancer therapy due to their selectivity, potency, and diverse mechanisms of action, their clinical application remains limited by several inherent challenges. Many ACPs suffer from short plasma half-lives, susceptibility to enzymatic degradation, poor bioavailability, and in some cases, unintended cytotoxicity to normal cells [35]. These limitations significantly impede the translational advancement of ACPs from preclinical research to clinical application. To overcome these challenges, researchers have increasingly focused on the chemical modification of ACPs, aiming to enhance their stability, specificity, and pharmacokinetic profiles while minimizing adverse effects. Broadly, the reconstruction of ACPs can be categorized into main chain modifications and side chain alterations. Main chain reconstruction typically involves the incorporation of non-natural amino acids, D-amino acids, or peptidomimetic backbones, which improve resistance to proteases and can stabilize desired secondary structures. On the other hand, side chain modifications include chemical additions such as cholesterol conjugation, phosphorylation, polyethylene glycol (PEG) attachment, glycosylation, and palmitoylation [7]. These strategies enhance membrane interaction, solubility, half-life, and targeted delivery. Besides the need for developing a comprehensive single-source access to different Anticancer peptides and proteins, there is a growing need for designing, analysing and predicting the chemically modified ACPs. Over the past decade numerous methods have been developed to classify natural ACPs. These include methods such as AntiCP and AntiCP2 which uses sequence information of the peptides and uses ML classifiers to perform binary classification. So far, a lot of ACP predictors using ML have been developed such as ACPP [36] , iACP [37] , iACP-GAEnsC [38], MLACP [17], AntiCP2.0 [16] and many more. Especially, ML models predict how peptide sequences affect target cells or diseases without physical and biological analyses, owing to advances in computer power, algorithm power utilizing datasets from dedicated Databases like CancerPPD [10]. Deep Learning is also used for ACP development in prediction methods such as ACP-DL [39] and PTPD [40]. These methods were limited to short natural peptides upto the length of 40. The introduction of AntiCP3 [41], which utilizes evolutionary information from protein language models (e.g., ESM2), marks a significant advancement by enabling the prediction of anticancer activity in longer protein sequences (50–1000 residues). All these methods did not address the issue of classifying chemically modified peptides. Despite substantial advancements in the development of computational models for natural anticancer peptides (ACPs), a critical gap persists: no existing

method to date enables the accurate prediction of anticancer or antimicrobial activity for chemically modified peptides, particularly those incorporating diverse non-canonical residues, backbone alterations, or terminal modifications. a significant proportion of FDA-approved peptide-based therapeutics are chemically modified. Such modifications are essential for enhancing proteolytic resistance, improving pharmacokinetic profiles, evading immune surveillance, and reducing off-target toxicity—attributes that are vital for therapeutic efficacy and clinical success. The absence of predictive frameworks tailored to this chemically diverse class of peptides constitutes a major bottleneck in the rational design and translational advancement of peptide therapeutics.

In the present study, we introduce the first systematic effort to predict chemically modified anticancer peptides by leveraging various types of **sequence**- and **SMILES**-based features. By explicitly accounting for the conformational and physicochemical properties of chemical modifications, our approach addresses a long-standing methodological void and offers a foundational framework for advancing chemically modified peptide therapeutics through computational design. Figure 4.1 below shows overall architecture of the study.

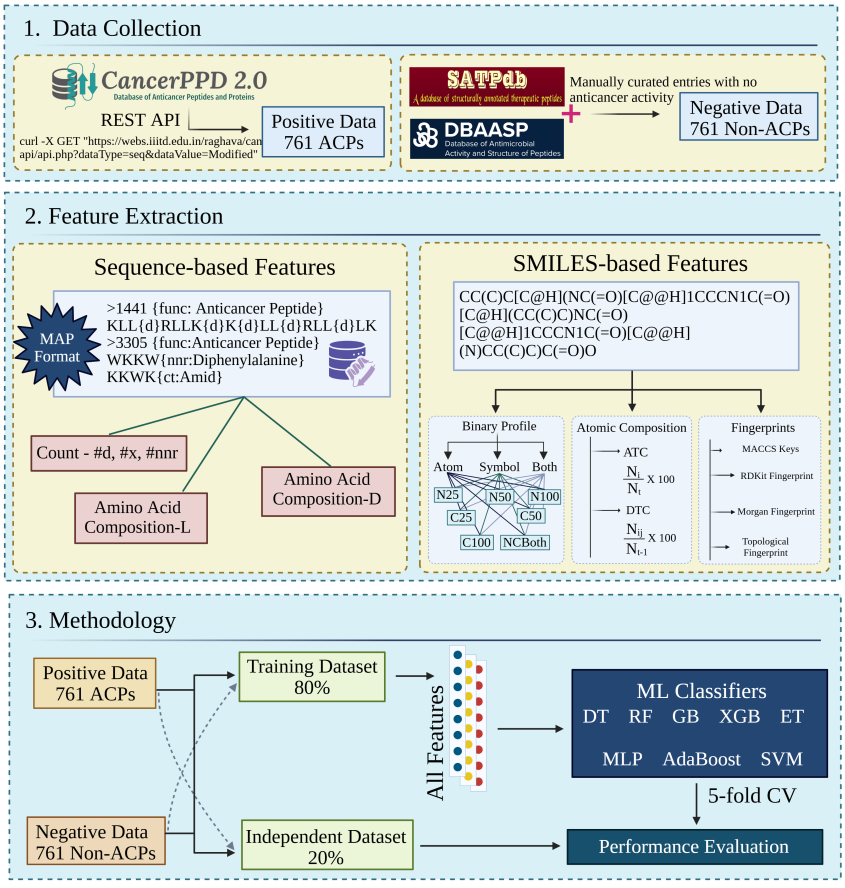


Fig. 4.1. *Overall architecture for the prediction of chemically modified anticancer peptides (ACPs).*

4.2 Materials Methods

4.2.1 Dataset Creation

To construct the **positive** dataset, we selected all peptides with any form of chemical modification, including N- or C-terminal modifications, incorporation of non-natural amino acids (e.g., D-amino acids), or other chemical alterations like non-natural residues such as ornithine, norleucine, homoarginine, etc. , provided that their sequence annotations were complete from CancerPPD2. The collected sequences were checked for redundancy finally resulting in a total of **761** chemically modified anticancer peptides (ACPs), which formed the positive class.

For the negative dataset, we curated peptides from DBAASP [42] and SATPDB [43] that possess antimicrobial activity but lack any reported anticancer activity. To ensure the absence of anticancer annotations, we conducted rigorous manual curation and cross-referencing. Given the biophysical similarity between antimicrobial and anticancer peptides, selecting AMPs with no documented anticancer effects provided a challenging yet biologically relevant negative class, enhancing the discriminative capacity of our predictive model. Finally, a balanced dataset comprising **761** chemically modified ACPs and **761** chemically modified AMPs was prepared to train and evaluate the classification framework.

4.2.2 Dataset Preprocessing

To ensure standardized representation of chemically modified peptides, all sequences were formatted in the **MAP** (Modification and Annotation in Proteins) format, developed by Raghava's group in 2025 [44]. This format enables rich, structured annotation of protein and peptide sequences, capturing both residue-specific modifications and global sequence-level metadata in a unified scheme. In MAP format, peptide sequences are expressed using the standard one-letter amino acid code, with inline curly-brace tags denoting specific chemical modifications at the residue level. Examples include phosphorylation, acetylation, methylation, cyclization, and the incorporation of non-natural or D-amino acids. Additionally, meta-information such as origin, activity type, or modification class is embedded in the sequence header, allowing for seamless integration with computational pipelines. This representation enables consistent and lossless encoding of complex modifications across diverse peptides, providing a computationally compatible input format for downstream structural modeling, feature extraction, and machine learning workflows. Importantly, it facilitates interoperability with existing tools for modified peptides and enhances reproducibility in dataset generation. Given below is a peptide entry of a chemically modified anticancer peptide taken from CancerPPD2.

>1330

`LHARE{d}IK{nnr:Orn}M{ct:Amid}`

4.2.3 Feature Generation

1. Sequence-based Features

The frequency of each of the 20 standard L-amino acids was calculated from the peptide sequences. D-amino acids were explicitly represented using the MAP format, where their one-letter codes were enclosed in curly braces (e.g., D for D-aspartic acid, K for D-lysine). This allowed the model to distinguish between L- and D-forms of the same amino acid. All other non-natural or chemically modified residues (e.g., ornithine, phosphoserine) were uniformly represented as X to avoid inflating the feature space with rare or ambiguous tokens. This standardized representation helped retain biologically meaningful information while ensuring compatibility with numerical encoding methods. In this way, different AAC-based feature matrices were generated:

1. AAC with Non-Canonical Aggregation (AAC+X, 21D): This feature extends traditional amino acid composition by integrating a single category for chemically modified and non-natural residues. Specifically, the frequencies of the 20 standard L-amino acids are calculated from the peptide sequence, and an additional 'X' feature aggregates all residues that do not belong to the canonical set. These include non-natural residues (e.g., *ornithine*, *norleucine*) and post-translational modifications that are otherwise difficult to represent explicitly.

2. Extended Stereochemical Composition with Dual Modification Encoding (AAC+X+x, 42D): This feature provides a granular view of the peptide sequence by capturing both L- and D-forms of amino acids, alongside modified residues. The feature vector comprises:

- Frequencies of 20 standard L-amino acids
- Frequencies of 20 D-amino acids, represented using MAP format (e.g., `{d}A`, `{d}L`, `{d}M`, etc.)
- A feature for non-natural/unknown residues in L-form, denoted as `X`
- A feature for non-natural/unknown residues in D-form, denoted as `x`

This 42-dimensional representation enables the model to differentiate between stereoisomers and interpret the biological relevance of stereospecific modifications, which can significantly influence peptide stability, protease resistance, and activity.

3. Composition with Explicit Modification Counts (AAC+ModCounts, 23D/43D): This feature incorporates quantitative descriptors for modifications alongside amino acid composition. Two variants are used depending on whether D-residues are counted in the composition:

- **23D:** 20 standard L-amino acids + counts of D-residues, counts of `X`, and counts of non-natural residues (NNR)

- **43D:** 40 AAC dimensions (20 L + 20 D amino acids) + counts of D-residues, counts of `X`, and counts of NNR

4. Composition with Structural and Modification Flags (AAC+ModFlags, 25D/45D): This feature integrates standard or extended amino acid composition with binary indicators representing biologically meaningful properties of peptides. Binary flags include:

- D-residue presence (1 if any D-aa, else 0)
- Non-natural residue (NNR) flag (1 if `X` or `x` present, else 0)
- Linear/Cyclic topology (1 for linear, 0 for cyclic)
- N-terminal modification (1 if present, else 0)
- C-terminal modification (1 if present, else 0)

These flags help encapsulate key biological and chemical contexts of the peptides without expanding the compositional space excessively.

2. SMILES - based Features

To incorporate chemically relevant information beyond amino acid sequences, peptide sequences were converted into **SMILES (Simplified Molecular Input Line Entry System)** representations. SMILES is a text-based notation that encodes the 2D molecular structure, capturing the atomic connectivity, bond types, formal charges, and stereochemistry of molecules. While peptide sequences convey residue-level information, SMILES representations describe the precise atomic-level arrangement of each peptide, including side chains, terminal modifications (e.g., *N-terminal acetylation, C-terminal amidation*), and stereochemistry such as **L**- and **D**-amino acid configurations. This level of detail allows the model to recognize subtle yet biologically significant variations in peptide structure that affect properties like binding affinity, membrane permeability, enzymatic stability, and overall bioactivity. For example, the inclusion of D-residues or chemical modifications can alter a peptide's resistance to proteolytic degradation or its ability to adopt a bioactive conformation — factors not directly captured by sequence alone.

Generation of SMILES

To streamline and standardize the generation of SMILES strings, we adopted a sequence-to-SMILES pipeline that bypasses the intermediate 3D structure (PDB) generation step. This was achieved using the **cyclicpeptide** Python package developed by Yang et al. [45], which provides utilities for constructing molecular structures directly from peptide sequences. Many chemically modified peptides in our dataset included non-standard monomers not present in the default monomer library of the tool (e.g., Cyclohexylalanine, ornithine,

phosphoserine, and terminal capping groups such as C-terminal amidation, N-terminal Palmitoylation, etc.). To address this, we curated and expanded the monomer library of the cyclicpeptide package by adding chemically accurate representations of dataset-specific monomers. These extensions ensured that all annotated modifications—including side-chain alterations, D-form residues, and terminal modifications—were faithfully incorporated during SMILES construction. This allowed us to preserve the full chemical integrity of both natural and chemically modified peptides during the SMILES generation process, ensuring fidelity in downstream structure-based feature extraction and modeling. Figure 4.2 illustrates the complete pipeline for generating SMILES representations from peptide sequences and extracting various features, which are discussed in detail in the subsequent sections.

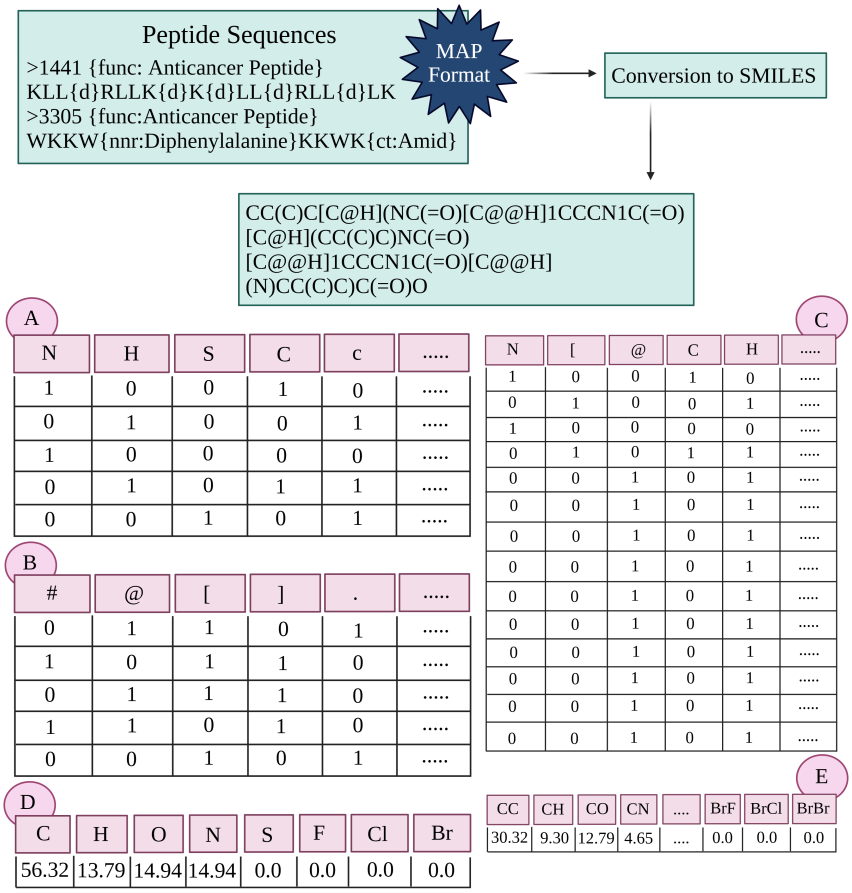


Fig. 4.2. *Feature extraction using SMILES format. Different features were calculated using SMILES format (A) binary profile generation of only atoms, (B) binary profile generation of only symbols, (C) binary profile generation of both symbol and atoms, (D) atom composition, and (E) diatom composition.*

1. Atom Composition: The Atomic composition (ATC) for modified ACPs non-ACPs was calculated using SMILES. The SMILES were further used to calculate atom composition of following atoms **C, H, O, N, S, Cl, Br, and**

F (see Figure 4.2). The atomic composition is calculated using formula 4.1 and provides a fixed length of eight vectors.

$$\text{Fraction of atom } (a) = \frac{\text{Total number of atom } (a)}{\text{Total number of all possible atoms}} \times 100 \qquad (4.1)$$

where *atom (a)* represents one out of all eight atom types considered.

2. Diatom Composition: Diatom composition was computed in a similar manner as atom composition. The diatomic composition provides information about the pairs of atoms in each residue (e.g., C-C, C-O, C-N, etc.) of the peptides (see Figure 4.2). The diatomic composition was computed using formula 4.2 which provided us a fixed length of **64 (8 × 8)** vectors.

$$\text{Fraction of atom } (a) = \frac{\text{Total number of diatom } (a)}{\text{Total number of all possible diatoms}} \times 100 \qquad (4.2)$$

where *diatom (a)* is one out of all 64 diatoms.

3. Binary Profile: Distinguishing anticancer peptides (ACPs) from non-ACPs with closely related sequences remains a challenging task in computational biology. While conventional features such as atomic composition and physicochemical descriptors provide valuable insights, they often fail to capture the positional context and sequential ordering of atoms—an important aspect of peptide bioactivity. To address this limitation, we used SMILES representation of peptides to retain chemical specificity and enable order-sensitive feature extraction.

From the SMILES strings, we extracted fixed-length segments of characters from the *N-terminus* and *-terminus* and generated binary profiles based on the occurrence of specific atoms and chemical symbols. We constructed predictive models under following feature categories:

i) **Atom-based profiles**, using only atomic elements.
For atom-based profiles, we focused on eight commonly occurring atoms in peptides—**C**, **H**, **O**, **N**, **S**, **F**, **Cl**, and **Br**. A binary vector was constructed where each atom's presence was encoded as '1' and absence as '0', resulting in a feature matrix of size $N \times 8$.

ii) **Symbol-based profiles**, using common SMILES notation characters.
For symbol-based profiles, we selected seven SMILES symbols: `@`, `+`, `=`, `#`, `[`, `]`, and `.`, which represent stereochemistry, charges, bond types, and structural annotations. A binary vector was constructed where each symbol's presence was encoded as '1' and absence as '0', resulting in a feature matrix of size $N \times 7$.

iii) **Combined Atom- and Symbol- based profiles**.
In this case, features from both atom-based and symbol-based profiles, as described above, were concatenated to form a single binary vector of

length $N \times 15$. This combined representation captures both elemental and structural features in the SMILES notation.

For each category, we considered sequence lengths of 50, 100, and 200 characters from the termini. The complete pipeline for binary profile generation is illustrated in Figure 4.2.

3. Chemical Descriptors: To characterize the molecular and structural properties of chemically modified anticancer peptides (ACPs), we utilized the **cyclicPeptide.PropertyAnalysis** module, a component of the CyclicPeptide Python package [45]. This module computes a wide range of physicochemical descriptors that are crucial for understanding drug-like behaviour, molecular complexity, and bioavailability of peptides. These descriptors provide essential features for downstream machine learning applications and comparative analyses.

The following categories of molecular descriptors were calculated:

- **Atomic and Structural Features**:
  - *Number of Atoms* – Total atom count, indicative of molecular size.
  - *Number of Rings* – Represents cyclic motifs within the peptide, often influencing rigidity and bioactivity.
  - *Heavy Atom Count* – Number of non-hydrogen atoms, reflecting molecular backbone complexity.
  - *Rotatable Bond Count* – Number of rotatable bonds, serving as a proxy for molecular flexibility.
  - *Formal Charge* – Net charge of the molecule, which can affect solubility and interaction with biological membranes.
- **Physicochemical Descriptors**:
  - *Exact Mass* – Monoisotopic molecular mass.
  - *Topological Polar Surface Area (TPSA)* – Summed surface area of polar atoms, predictive of intestinal absorption and blood–brain barrier penetration.
  - *Crippen LogP* – Logarithm of the octanol–water partition coefficient, a measure of hydrophobicity.
  - *Complexity* – A quantification of structural intricacy, considering symmetry, atom types, and bonding patterns.
  - *Refractivity* – Reflects molecular polarizability and electronic interactions.
- **Hydrogen Bonding Capacity**:
  - *Hydrogen Bond Donor Count* – Number of hydrogen bond donors (typically NH or OH groups), relevant for target binding.

- *Hydrogen Bond Acceptor Count* – Number of acceptors (typically N or O atoms), critical for ligand–receptor interactions.

- **Molecular Fingerprints**:
  - *RDKit Fingerprint* – A hashed molecular fingerprint representing molecular substructures using RDKit's default settings.
  - *Morgan Fingerprint* – A circular fingerprint (similar to ECFP), encoding substructural environments around each atom using functional group features.
  - *MACCS Keys* – A fixed set of SMARTS-based substructure keys, often used in virtual screening and QSAR modelling.
  - *Daylight-like Fingerprint* – A fingerprint mimicking Daylight methodology, capturing connectivity-based paths in the molecule.

In total, this approach yielded approximately **5301** features per peptide, offering a comprehensive and chemically meaningful vector representation of each sequence.

4.2.4 Compositional Analysis

The average atomic composition of chemically modified ACPs and non-ACPs was calculated to evaluate atom-type preferences between the two groups. A two-sample independent t-test was conducted to assess the statistical significance of these differences, with a p-value < 0.05 considered indicative of statistical significance.

4.2.5 Model Development

ML classifiers that are based on data-driven learning are widely used in protein/peptide classification tasks. In this study, we have developed various classification models for binary classification of modified Anticancer peptides. We have used models like Tree-based classifiers such as Decision Tree (DT), Random Forest (RF) and Extra Tree (ET), Ensemble methods with boosting strategies like AdaBoost (AB), Gradient Boosting (GB) and Extreme Gradient Boost (XGB), linear classifier like Logistic Regression (LR), lazy learners like K-nearest neigh bour (KNN), kernel based classifiers like Support Vector Classifier (SVC), and Neural-network based model like Multi-layer Perceptron classifier (MLP) and optimized them using various hyperparameters best suited for our dataset.

4.2.6 Cross Validation Performance evaluation

The complete dataset was stratified into two subsets: (i) a training (or internal) dataset and (ii) a validation (or external) dataset, following the standard practices in ML training algorithms. The training dataset comprised 80% of

the total data, including 608 chemically modified ACPs and 609 non-ACPs, selected at random to mitigate any sampling bias. The remaining 20% of the data, consisting of 153 modified ACPs and 152 non-ACPs, was reserved for external validation. Model development and hyperparameter tuning were conducted on the training dataset using a fivefold cross-validation approach, ensuring robust internal evaluation. The final optimized models were subsequently assessed on the independent validation dataset to evaluate their generalizability and real-world applicability.

To comprehensively assess the performance of the developed models, following evaluation metrices were used to evaluate the performance -

$$\text{Sensitivity} = \frac{TP}{TP + FN} \times 100$$

$$\text{Specificity} = \frac{TN}{TN + FP} \times 100$$

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \times 100$$

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

4.3 Results

4.3.1 Compositional Analysis

Atom Composition

The percent average composition of atoms present in modified ACPs and non-ACPs was computed to understand atom-level preferences. Both groups exhibit high proportions of hydrogen (54%) and carbon (31%), reflecting the predominance of hydrocarbon backbones in peptide structures, which form the core of both side chains and main chains. Notably, oxygen content is significantly lower in ACPs than in non-ACPs. Given that oxygen-rich groups such as carboxyl and hydroxyl are often associated with solubility and polarity, this reduction may suggest differences in hydrophilicity or functional group diversity between the two classes. Nitrogen content also varies significantly, indicating possible differences in the presence of functional groups like amides, amines, or guanidinium moieties—commonly found in positively charged residues such as lysine and arginine, which are known to facilitate membrane interaction and contribute to anticancer activity. Although present in very low percentages, atoms like fluorine and sulfur are statistically enriched in ACPs. Fluorine may result from halogenation strategies or the inclusion of fluorinated amino acids, potentially enhancing peptide stability or membrane affinity. Sulfur likely reflects the incorporation of sulfur-containing residues such as cysteine or methionine, which can influence disulfide bonding and structural rigidity. In contrast,

chlorine and bromine show negligible contributions in both peptide classes, suggesting halogenation is not a common natural modification. Overall, the observed differences in nitrogen, oxygen, fluorine, and sulfur composition highlight underlying chemical modifications and residue biases that may influence properties such as bioavailability, membrane interaction, and reactivity. Figure 4.3 shows the comparative atomic composition in the two groups highlighting the residues with statistically significant differences.

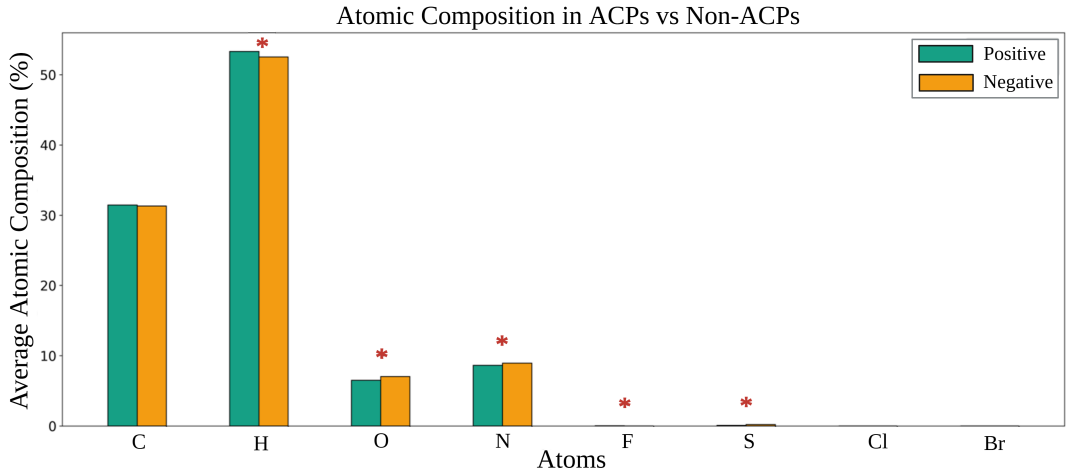


Fig. 4.3. *Average Atomic Composition in chemically modified ACPs(Positive) and non-ACPs(negative)*

Di-Atom Composition

The graph in the Figure 4.4 below depicts the average diatomic atomic composition (% of all atom pairs) of molecules in anticancer peptides (ACPs) versus non-anticancer peptides (non-ACPs), calculated from their SMILES representations. Diatomic composition (DTC) provides insights into local atomic connectivity patterns, which can be more informative than single-atom statistics for understanding molecular structure–activity relationships. The C–H and C–C pairs are the most prevalent diatomic pairs across both classes, consistent with their central role in defining the hydrocarbon backbone of peptides. Despite similar dominance, both C–C and C–H show statistically significant differences (p ¡ 0.05), with ACPs showing slightly higher proportions. This suggests potential enrichment of aliphatic or saturated chain structures in ACPs. C–N and C–O bonds, which are associated with peptide bonds, amines, and side-chain functionalities, show significantly different proportions between ACPs and non-ACPs. A higher proportion of C–N in non-ACPs might indicate a tendency toward more nitrogen-rich motifs outside the anticancer space, whereas C–O bonds are slightly higher in ACPs, potentially linked to enhanced hydrogen bonding or oxygenated functional groups beneficial for anticancer activity. Atom pairs such as C–S, H–O, and H–S occur at very low frequencies, but the differences remain statistically significant. These subtle variations might correspond to sulfur- or hydroxyl-containing residues like cysteine, methionine,

serine, or tyrosine, which are known to modulate bioactivity and structural stability in bioactive peptides.

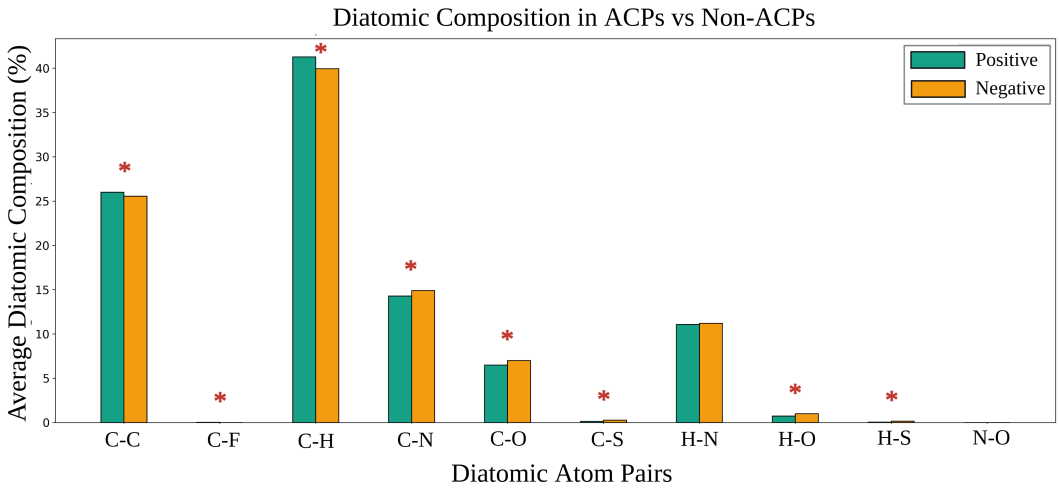


Fig. 4.4. *Average Di-Atomic Composition in chemically modified ACPs(Positive) and non-ACPs(negative)*

4.3.2 ML based Prediction Model

Sequence-based Features

ML classifiers as described above were used for this task of binary classification of chemically modified ACPs non-ACPs. We used different representations of sequence-based features to evaluate their effectiveness. We calculated the Amino Acid Composition (AAC) based solely on the 20 standard L-amino acids, resulting in a 20-dimensional (20D) feature vector. Among the models tested ET-based classifiers performed best with an AUC score of **0.89** on validation set. These features were not the true representations of chemically modified peptides inadequately captured the chemical complexity and diversity introduced by peptide modifications. To address this limitation, we extended the AAC representation to include D-amino acids and non-natural residues, and tried various combinations. Detailed results are shown in Table 4.1 below.

Feature Type	**Model**	Sens	Spec	Acc	AUC	MCC
AAC (20)	Extra Trees	0.81	0.80	0.81	0.89	0.61
AAC + X (21)	Random Forest	0.79	0.77	0.78	0.88	0.57
AAC + Count of D + X + NNR (23)	Extra Trees	0.78	0.77	0.77	0.85	0.55
AAC + %Composition of D + X + NNR (43)	Random Forest	0.77	0.77	0.77	0.83	0.53
AAC + Mod Flags (25)	Extra Trees	0.77	0.77	0.77	0.84	0.54
AAC + Mod Flags (46)	Extra Trees	0.75	0.76	0.76	0.84	0.51

Table 4.1. *Performance of different models using various combinations of sequence-based features.*

SMILES-based Features

i. Composition-based Features

We developed prediction models for the atomic and diatomic composition of the peptide using various classifiers. In case of atomic composition, ET-based classifier performed better than other classifiers with an AUC score of **0.74** and RF performed better in case of DTC with an AUC score of **0.78** on validation dataset. Detailed Results for ATC (see Table 4.2) and DTC (see Table 4.3) are given below.

Model	**Training**					**Validation**				
	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
Decision Tree	0.64 ± 0.03	0.63 ± 0.08	0.64 ± 0.04	0.64 ± 0.04	0.27 ± 0.08	0.62	0.61	0.61	0.61	0.23
Random Forest	0.69 ± 0.02	0.68 ± 0.01	0.69 ± 0.02	0.75 ± 0.03	0.37 ± 0.03	0.66	0.66	0.66	0.73	0.32
Gradient Boosting	0.55 ± 0.04	0.68 ± 0.06	0.62 ± 0.03	0.69 ± 0.03	0.23 ± 0.07	0.44	0.70	0.57	0.62	0.14
AdaBoost	0.65 ± 0.03	0.65 ± 0.02	0.65 ± 0.03	0.72 ± 0.03	0.30 ± 0.05	0.62	0.60	0.61	0.66	0.22
XGBoost	0.70 ± 0.04	0.66 ± 0.04	0.66 ± 0.04	0.72 ± 0.03	0.32 ± 0.08	0.60	0.61	0.61	0.68	0.21
Extra Trees	0.68 ± 0.03	0.67 ± 0.03	0.68 ± 0.03	0.75 ± 0.03	0.35 ± 0.06	0.66	0.67	0.67	0.74	0.33
Logistic Regression	0.59 ± 0.02	0.59 ± 0.04	0.59 ± 0.03	0.64 ± 0.01	0.18 ± 0.05	0.60	0.57	0.58	0.62	0.17
KNN	0.68 ± 0.06	0.64 ± 0.06	0.66 ± 0.03	0.70 ± 0.04	0.31 ± 0.06	0.61	0.63	0.62	0.69	0.24
SVC	0.63 ± 0.03	0.62 ± 0.03	0.62 ± 0.03	0.69 ± 0.03	0.25 ± 0.06	0.59	0.59	0.59	0.64	0.18
MLP	0.66 ± 0.04	0.66 ± 0.04	0.66 ± 0.04	0.72 ± 0.03	0.32 ± 0.08	0.60	0.59	0.60	0.66	0.20

Table 4.2. *The table shows the performance of various machine learning classifiers on Atomic Composition (ATC)*

Model	Training					Validation				
	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
Decision Tree	0.70 ± 0.04	0.67 ± 0.03	0.680 ± 0.003	0.69 ± 0.04	0.37 ± 0.07	0.66	0.66	0.66	0.65	0.32
Random Forest	0.74 ± 0.01	0.73 ± 0.02	0.732 ± 0.002	0.80 ± 0.01	0.46 ± 0.03	0.69	0.70	0.70	0.78	0.39
Gradient Boosting	0.68 ± 0.03	0.65 ± 0.04	0.670 ± 0.003	0.72 ± 0.02	0.33 ± 0.05	0.56	0.68	0.62	0.65	0.24
AdaBoost	0.71 ± 0.02	0.71 ± 0.02	0.710 ± 0.002	0.77 ± 0.02	0.42 ± 0.04	0.70	0.70	0.70	0.73	0.39
XGBoost	0.71 ± 0.03	0.71 ± 0.03	0.710 ± 0.003	0.78 ± 0.02	0.42 ± 0.05	0.68	0.68	0.68	0.77	0.36
Extra Trees	0.73 ± 0.02	0.73 ± 0.02	0.730 ± 0.002	0.80 ± 0.02	0.45 ± 0.03	0.68	0.67	0.68	0.77	0.35
Logistic Regression	0.65 ± 0.03	0.62 ± 0.03	0.630 ± 0.003	0.69 ± 0.02	0.26 ± 0.05	0.61	0.64	0.63	0.66	0.25
KNN	0.72 ± 0.04	0.65 ± 0.03	0.680 ± 0.003	0.76 ± 0.03	0.37 ± 0.07	0.64	0.67	0.66	0.70	0.32
SVC	0.66 ± 0.01	0.66 ± 0.01	0.660 ± 0.001	0.72 ± 0.02	0.32 ± 0.02	0.60	0.62	0.61	0.65	0.22
MLP	0.71 ± 0.03	0.71 ± 0.03	0.710 ± 0.003	0.76 ± 0.02	0.42 ± 0.06	0.66	0.67	0.67	0.72	0.33

Table 4.3. *The table shows the performance of various machine learning classifiers on Di-atomic Composition (DTC)*

ii. Binary Profiles

Binary profiles were generated using the `SMILES` representations of chemically modified peptides. Two categories of features were considered.

First, we focused on **atomic-level representations**. Only atom symbols (e.g., C, N, O, S, etc.) were considered, and binary vectors were constructed based on the presence or absence of atoms at specific positions. Profiles were generated by extracting the first 25, 50, and 100 atoms from:

- the N-terminal end (`N25`, `N50`, `N100`),
- the C-terminal end (`C25`, `C50`, `C100`), and
- a combination of both terminals (`N25C25`, `N50C50`, `N100C100`).

The binary vectors were built using one-hot encoding for each atom at each position. Among these, the best classification performance was obtained using `N50C50`, which achieved an AUC of **0.80** on the validation dataset.

Second, we considered only **non-alphabetic SMILES symbols** such as `@, +, =, #, [, ], .` to capture stereochemical and structural information. Binary profiles were similarly created for the first 25, 50, and 100 symbols from the N-terminal, C-terminal, and both terminals (e.g., `symbol_N25`, `symbol_C50`, `symbol_N50C50`). The highest performance in this category was obtained using `N50C50` and also using only `N50`, both achieving an AUC of **0.80**.

Furthermore, we integrated both atom-based and symbol-based binary profiles by extracting elements from the N-terminus, C-terminus, and a combination of both termini. Among all combinations, the combined `Both_N100C100`, achieved highest predictive performance with an AUC of **0.83** on the validation set. These binary profiles allowed for capturing structural and stereochemical characteristics from SMILES strings and provided complementary information to traditional sequence-based features.

Detailed performance for all combinations is given in Table 4.4 below.

Feature Type	**Model Name**	Sens	Spec	Acc	AUC	MCC
Atom N25	Extra Trees	0.70	0.68	0.69	0.75	0.38
Atom N50	Extra Trees	0.71	0.70	0.70	0.77	0.41
Atom N100	Random Forest	0.69	0.69	0.69	0.77	0.38
Atom C25	Decision Tree	0.70	0.73	0.71	0.72	0.43
Atom C50	Random Forest	0.69	0.71	0.70	0.78	0.40
Atom C100	Random Forest	0.66	0.65	0.66	0.76	0.66
Atom N25C25	Random Forest	0.72	0.70	0.71	0.79	0.42
Atom N50C50	Extra Trees	0.70	0.71	0.71	0.80	0.41
Atom N100C100	Extra Trees	0.72	0.67	0.70	0.78	0.39
Symbol N25	XGBoost	0.71	0.71	0.71	0.80	0.42
Symbol C25	Extra Trees	0.73	0.68	0.70	0.72	0.41
Symbol N50	Random Forest	0.73	0.73	0.73	0.80	0.45
Symbol C50	XGBoost	0.70	0.69	0.69	0.76	0.38
Symbol N100	Random Forest	0.68	0.70	0.69	0.77	0.38
Symbol C100	Extra Trees	0.69	0.70	0.69	0.75	0.38
Symbol N25C25	MLP	0.74	0.73	0.74	0.79	0.47
Symbol N50C50	XGBoost	0.71	0.71	0.71	0.80	0.42
Symbol N100C100	Extra Trees	0.69	0.67	0.68	0.76	0.36
Both N25	Random Forest	0.75	0.73	0.74	0.81	0.48
Both C25	Random Forest	0.71	0.71	0.71	0.79	0.42
Both N25C25	XGBoost	0.68	0.65	0.67	0.74	0.33
Both N50	Random Forest	0.76	0.75	0.75	0.79	0.50
Both C50	Random Forest	0.72	0.75	0.73	0.81	0.46
Both N50C50	Extra Trees	0.71	0.69	0.70	0.79	0.40
Both N100	Extra Trees	0.71	0.73	0.72	0.79	0.44
Both C100	Extra Trees	0.70	0.72	0.71	0.77	0.41
Both N100C100	Extra Trees	0.75	0.73	0.74	0.83	0.48

Table 4.4. *Performance of various machine learning classifiers on different Binary Profiles generated using SMILES over validation dataset*

iii. Chemical Descriptors

As discussed in the previous section, the **cyclicpeptide** package was used to compute a wide range of chemical descriptors directly from the SMILES representations of peptides. These descriptors captured detailed structural, physicochemical, and electronic properties of the molecules, including features such as topological polar surface area, molecular complexity, hydrogen bond donors/acceptors, formal charge, rotatable bond count, and multiple types of molecular fingerprints (e.g., Morgan, MACCS, RDKit, etc.). Among the various machine learning classifiers evaluated, the Extra Trees model achieved the highest classification performance with an AUC score of **0.88** on the validation

dataset. Detailed results are shown in Table 4.5 below.

Models	Training					Validation				
	Sens	Spec	Acc	AUC	MCC	Sens	Spec	Acc	AUC	MCC
Decision Tree	0.72 ± 0.05	0.71 ± 0.02	0.71 ± 0.02	0.72 ± 0.02	0.43 ± 0.04	0.64	0.67	0.66	0.66	0.32
Random Forest	0.78 ± 0.02	0.78 ± 0.03	0.78 ± 0.02	0.85 ± 0.03	0.57 ± 0.05	0.76	0.77	0.77	0.84	0.53
Gradient Boosting	0.68 ± 0.02	0.68 ± 0.03	0.68 ± 0.01	0.76 ± 0.01	0.36 ± 0.03	0.70	0.70	0.70	0.76	0.40
AdaBoost	0.74 ± 0.03	0.75 ± 0.03	0.74 ± 0.03	0.82 ± 0.01	0.49 ± 0.06	0.76	0.76	0.76	0.81	0.52
XGBoost	0.78 ± 0.02	0.78 ± 0.02	0.78 ± 0.02	0.87 ± 0.02	0.56 ± 0.05	0.77	0.77	0.77	0.85	0.54
Extra Trees	0.77 ± 0.02	0.77 ± 0.02	0.77 ± 0.02	0.84 ± 0.03	0.54 ± 0.04	0.80	0.79	0.80	0.88	0.59
Logistic Regression	0.73 ± 0.03	0.74 ± 0.03	0.73 ± 0.03	0.79 ± 0.03	0.47 ± 0.06	0.71	0.71	0.71	0.77	0.42
KNN	0.73 ± 0.03	0.68 ± 0.05	0.71 ± 0.03	0.76 ± 0.04	0.41 ± 0.06	0.64	0.70	0.67	0.77	0.35
SVC	0.73 ± 0.02	0.73 ± 0.02	0.73 ± 0.02	0.81 ± 0.02	0.46 ± 0.04	0.73	0.73	0.73	0.80	0.47
MLP	0.77 ± 0.03	0.77 ± 0.03	0.77 ± 0.03	0.83 ± 0.04	0.53 ± 0.05	0.76	0.75	0.76	0.82	0.51

Table 4.5. *Performance comparison of different models on training and validation sets for Chemical Descriptors.*

These results demonstrate the effectiveness of using chemically rich descriptors derived from SMILES strings in capturing subtle variations introduced by non-natural modifications in peptides, thereby enhancing the model's ability to discriminate between anticancer and non-anticancer peptides.

4.4 Discussion

Anticancer peptides (ACPs) have garnered significant attention in recent years owing to their selectivity, rapid action, and ability to evade traditional mechanisms of drug resistance. Anticancer peptides (ACPs) exhibit considerable heterogeneity in their physicochemical characteristics, amino acid composition, structural conformations, and mechanisms of action [7]. Most ACPs are relatively short (ranging from 5 to 50 amino acids), cationic in nature, and display amphipathic properties, which facilitate their preferential interaction with the negatively charged membranes of cancer cells. Their diverse primary sequences include both hydrophobic and polar residues in varying proportions, enabling distinct structural motifs such as -helices, -sheets, extended structures, and cyclic conformations [7]. This structural versatility translates into a broad spectrum of anticancer mechanisms, including membrane disruption, mitochondrial destabilization, inhibition of angiogenesis, and immune modulation. Moreover, ACPs can selectively target tumor cells while sparing normal cells, largely due to differences in membrane composition, electrostatics, and fluidity — a feature that sets them apart from conventional chemotherapeutics.

The natural ecosystem serves as a prolific reservoir of bioactive compounds, including a rich repertoire of peptides with therapeutic potential. Numerous antimicrobial and anticancer peptides (AMPs/ACPs) have been isolated from diverse biomes, spanning microorganisms, plants, and animals. Notable examples include bovine *lactoferrin* and *LL-37* [46, 47], which are higly potent against Cancer cell lines and have been extensively studied for their therapeutic roles. Despite the availability of different literature evidences where these

peptides are shown to have high anticancer activity against different Cancer cell lines only a limited number of natural ACPs have advanced to clinical evaluation due to inherent challenges such as proteolytic instability, rapid clearance, and poor bioavailability. To address these limitations, chemical modifications — including the incorporation of D-amino acids, N-terminal and C-terminal capping, cyclization, and the use of non-natural residues — have emerged as critical strategies to enhance the therapeutic efficacy and pharmacokinetic profile of ACPs. For e.g. CancerPPD2 hosts different synthetic analogues of LL-37 a natural Antimicrobial peptide with different synthetic modifications like N-terminal Acetylation and C-terminal Amidations that have significantly improve their Anticancer activity.

Despite the growing relevance of these chemically modified peptides, computational efforts to predict or model their anticancer potential remain scarce. Most existing methods and databases have been developed and optimized primarily for natural L-peptides, with limited regard for stereoisomerism, non-canonical residues, or structural modifications [7]. This represents a significant bottleneck in the field, as chemical modifications are not peripheral alterations — they fundamentally reshape the peptide's conformation, charge distribution, hydrophobicity, and interaction landscape with cancer cell membranes or intracellular targets. Different online tools are available such as `AntimpMOD` [48] and `HemopiMOD` [49] that can predict chemically modified antimicrobial peptides and hemolytic peptides, but no such tools are available which are dedicated for Anticancer Peptides. In this study we have made an initial attempt to classify chemically modified ACPs.

We have used different features to study different aspects of chemically modified peptides. Sequence based features may seem to give very high results but they can be misleading as not all the information were correctly encoded. We have tried to extend these features by adding binary flags and representation of D-amino acids, but due to such large library of chemically modified peptides incorporation of each and every non-natural residue was not possible. One of the crucial aspect to study peptides is through their 3D structures, but due to limited number of force fields not all the modified peptides could be predicted. Although tools such as `PepstrMOD` is an state-of-the-art methods to predict tertiary structures of modified peptides, it was not possible to predict the structure for all the modifications. Also, to bypass this step of structure prediction we have tried to generate the SMILES directly from the sequences by representing modifications in a standard computer-readable format MAP format using cyclicpeptide package. Features generated from these SMILES showed promising results. Among them the chemical descriptors based features gave a high AUC score of 0.88 showing that they were able to capture the structural and chemical properties of modified peptides, including stereochemistry, electronic effects, and functional group modifications. These findings underscore the importance of chemically-aware representations and set the foundation for future models that integrate sequence, structure, and chemical context to enhance prediction accuracy and biological relevance in peptide-based therapeutic

discovery.

4.5 Limitations & Future Work

While this study provides a foundational effort toward the classification of chemically modified anticancer peptides (ACPs), several important limitations must be acknowledged. Firstly, although the use of SMILES enabled the extraction of atom-level structural features, our approach relied on the direct generation of SMILES from annotated peptide sequences, rather than from experimentally resolved or computationally predicted 3D structures. This was necessitated by the lack of high-quality 3D structures for many modified peptides, particularly those containing non-natural or synthetically derived residues, which are poorly represented in current structural databases. The absence of such structural data limits the accuracy with which conformational and spatial features—critical for peptide–target interactions—can be captured. To address this, we extended the monomer library of the cyclicpeptide package to include dataset-specific modifications such as C-terminal amidation and cyclohexyl alanine. While this allowed for a broader representation of chemically diverse peptides, the system may still fail to accurately model rare or uncharacterized modifications, thus limiting its generalizability to peptides beyond the training distribution. Secondly, the study focused exclusively on binary classification (ACP vs. non-ACP) and does not attempt to quantify or predict the anticancer potency or mechanism of action of the peptides. Future studies should consider integrating quantitative bioactivity data or multi-label classification frameworks. Finally, the lack of three-dimensional structural integration is a critical limitation. Structural features such as peptide folding, backbone conformation, and interaction interfaces significantly influence biological activity. However, the absence of reliable 3D models—owing to limited availability of force fields and parameter sets for modified residues—precluded their incorporation in this work. The development of robust structure prediction tools and chemically informed force fields will be essential to overcoming this barrier in future investigations.

We aim to develop a user-friendly web server to facilitate the broader scientific community in predicting and designing chemically modified anticancer peptides (ACPs). This platform will allow researchers to input peptide sequences—potentially including D-amino acids and non-natural residues—and receive predictions on their anticancer potential based on the trained models developed in this study. The web server will also incorporate tools for visualizing amino acid composition, identifying structural or chemical modification flags, and generating standardized SMILES representations. By providing an accessible interface and supporting customized peptide design, this resource will bridge the gap between computational predictions and experimental validation, accelerating the discovery of novel and therapeutically viable modified peptides.

Chapter 5

Summary

Cancer remains one of the most challenging diseases globally, demanding the continuous development of novel therapeutic strategies. Peptide- and Protein-based therapeutics are gaining attention due to lesser side effects than currently available conventional treatment strategies. While numerous studies have explored natural ACPs, chemically modified peptides—which often demonstrate enhanced stability, membrane permeability, and protease resistance—remain underexplored due to the complexity of their structures and the lack of comprehensive computational resources.

The integration of curated databases and robust in silico prediction tools has the potential to substantially accelerate the pipeline from peptide discovery to clinical trials. Manually curated datasets of chemically modified ACPs serve as a foundational resource by providing standardized, biologically validated information on peptide sequences, structural modifications, and anticancer activity. This curated knowledge base reduces experimental redundancy and supports rational peptide design by identifying chemical patterns and modifications associated with therapeutic efficacy. Furthermore, machine learning-based prediction models trained on these curated datasets enable rapid screening of large peptide libraries, prioritizing candidates with high anticancer potential and favorable biochemical properties. This in silico pre-screening can significantly reduce the time, cost, and resource burden of early-stage experimental validation, allowing researchers to focus on high-confidence candidates for in vitro and in vivo testing. Ultimately, such computational frameworks serve as valuable decision-support tools, guiding the selection of lead compounds with enhanced stability, bioavailability, and tumor selectivity—key attributes for successful translation into clinical trials.

We believe that the development of an updated repository of anticancer peptides—`CancerPPD2` - will play a pivotal role in advancing peptide-based therapeutics. By incorporating manually curated chemically modified ACPs along with their structural annotations, activity profiles, and experimental evidence, `CancerPPD2` aims to provide a comprehensive, high-fidelity resource for the research community. Our study also highlights the need for prediction tools for predicting chemically-modified ACPs, as this domain remainslargely unexplored. In conclusion, our integrated platform combining curated data and predictive modelling holds significant potential to accelerate peptide-based drug discovery. By enabling in silico screening and rational modification of ACPs, it paves the way for cost-effective and targeted therapeutic development in oncology.

References

[1] J. S. Brown, S. R. Amend, R. H. Austin, R. A. Gatenby, E. U. Hammarlund, and K. J. Pienta, "Updating the Definition of Cancer," *Molecular Cancer Research*, vol. 21, pp. 1142–1147, Nov. 2023.

[2] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, pp. 583–589, Aug. 2021.

[3] G. Ghaly, H. Tallima, E. Dabbish, N. Badr ElDin, M. K. Abd El-Rahman, M. A. A. Ibrahim, and T. Shoeib, "Anti-cancer peptides: Status and future prospects," *Molecules*, vol. 28, no. 3, 2023.

[4] S. M. P. Vadevoo, S. Gurung, H.-S. Lee, G. R. Gunassekaran, S.-M. Lee, J.-W. Yoon, Y.-K. Lee, and B. Lee, "Peptides as multifunctional players in cancer therapy," *Experimental & Molecular Medicine*, vol. 55, pp. 1099–1109, June 2023.

[5] L. Otvos, "Peptide-Based Drug Design: Here and Now," in *Peptide-Based Drug Design* (L. Otvos, ed.), pp. 1–8, Totowa, NJ: Humana Press, 2008.

[6] M. Bidram and M. R. Ganjalikhany, "Bioactive peptides from food science to pharmaceutical industries: Their mechanism of action, potential role in cancer treatment and available resources," *Heliyon*, vol. 10, p. e40563, Dec. 2024.

[7] M. Xie, D. Liu, and Y. Yang, "Anti-cancer peptides: classification, mechanism of action, reconstruction and modification," *Open Biology*, vol. 10, p. 200004, July 2020. Publisher: Royal Society.

[8] R. K. Chinnadurai, N. Khan, G. K. Meghwanshi, S. Ponne, M. Althobiti, and R. Kumar, "Current research status of anti-cancer peptides: Mechanism of action, production, and clinical applications," *Biomedicine & Pharmacotherapy*, vol. 164, p. 114996, 2023.

[9] S. Jain, S. Gupta, S. Patiyal, and G. P. Raghava, "Thpdb2: compilation of fda approved therapeutic peptides and proteins," *Drug Discovery Today*, p. 104047, 2024.

[10] A. Tyagi, A. Tuknait, P. Anand, S. Gupta, M. Sharma, D. Mathur, A. Joshi, S. Singh, A. Gautam, and G. P. Raghava, "CancerPPD: a

database of anticancer peptides and proteins," *Nucleic Acids Research*, vol. 43, pp. D837–D843, Jan. 2015.

[11] A. Tyagi, P. Kapoor, R. Kumar, K. Chaudhary, A. Gautam, and G. P. S. Raghava, "In Silico Models for Designing and Discovering Novel Anticancer Peptides," *Scientific Reports*, vol. 3, p. 2984, Oct. 2013.

[12] X. Xu, C. Li, X. Yuan, Q. Zhang, Y. Liu, Y. Zhu, and T. Chen, "ACP-DRL: an anticancer peptides recognition method based on deep representation learning," *Frontiers in Genetics*, vol. 15, Apr. 2024. Publisher: Frontiers.

[13] X. Liang, H. Zhao, and J. Wang, "Ma-pep: A novel anticancer peptide prediction framework with multimodal feature fusion based on attention mechanism," *Protein Science*, vol. 33, no. 4, p. e4966, 2024.

[14] J. Bian, X. Liu, G. Dong, C. Hou, S. Huang, and D. Zhang, "ACP-ML: A sequence-based method for anticancer peptide prediction," *Computers in Biology and Medicine*, vol. 170, p. 108063, Mar. 2024.

[15] M. Liu, T. Wu, X. Li, Y. Zhu, S. Chen, J. Huang, F. Zhou, and H. Liu, "ACPPfel: Explainable deep ensemble learning for anticancer peptides prediction based on feature optimization," *Frontiers in Genetics*, vol. 15, Feb. 2024. Publisher: Frontiers.

[16] P. Agrawal, D. Bhagat, M. Mahalwal, N. Sharma, and G. P. S. Raghava, "AntiCP 2.0: an updated model for predicting anticancer peptides," *Briefings in Bioinformatics*, vol. 22, p. bbaa153, May 2021.

[17] L. Thi Phan, H. Woo Park, T. Pitti, T. Madhavan, Y.-J. Jeon, and B. Manavalan, "MLACP 2.0: An updated machine learning tool for anticancer peptide prediction," *Computational and Structural Biotechnology Journal*, vol. 20, pp. 4473–4480, Jan. 2022.

[18] S. Burley, R. Bhatt, C. Bhikadiya, C. Bi, A. Biester, P. Biswas, S. Bittrich, S. Blaumann, R. Brown, H. Chao, V. R. Chithari, P. Craig, G. Crichlow, J. Duarte, S. Dutta, Z. Feng, J. Flatt, S. Ghosh, D. Goodsell, R. K. Green, V. Guranovic, J. Henry, B. Hudson, M. Joy, J. Kaelber, I. Khokhriakov, J.-S. Lai, C. Lawson, Y. Liang, D. Myers-Turnbull, E. Peisach, I. Persikova, D. Piehl, A. Pingale, Y. Rose, J. Sagendorf, A. Sali, J. Segura, M. Sekharan, C. Shao, J. Smith, M. Trumbull, B. Vallat, M. Voigt, B. Webb, S. Whetstone, A. Wu-Wu, T. Xing, J. Young, A. Zalevsky, and C. Zardecki, "Updated resources for exploring experimentally-determined pdb structures and computed structure models at the rcsb protein data bank," *Nucleic Acids Research*, vol. 53, pp. D564–D574, 11 2024.

[19] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Žídek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy,

M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, "Highly accurate protein structure prediction with AlphaFold," *Nature*, vol. 596, pp. 583–589, Aug. 2021.

[20] S. Singh, H. Singh, A. Tuknait, K. Chaudhary, B. Singh, S. Kumaran, and G. P. S. Raghava, "PEPstrMOD: structure prediction of peptides containing natural, non-natural and modified residues," *Biology Direct*, vol. 10, p. 73, Dec. 2015.

[21] H. Kaur, A. Garg, and G. P. S. Raghava, "Pepstr: a de novo method for tertiary structure prediction of small bioactive peptides," *Protein and peptide letters*, vol. 14, no. 7, pp. 626–631, 2007.

[22] W. Zheng, C. Zhang, Y. Li, R. Pearce, E. W. Bell, and Y. Zhang, "Folding non-homologous proteins by coupling deep-learning contact maps with i-tasser assembly simulations," *Cell reports methods*, vol. 1, no. 3, 2021.

[23] A. S. Rose, A. R. Bradley, Y. Valasatava, J. M. Duarte, A. Prlić, and P. W. Rose, "Web-based molecular graphics for large complexes," in *Proceedings of the 21st international conference on Web3D technology*, pp. 185–186, 2016.

[24] T. Madden, "The blast sequence analysis tool," *The NCBI handbook*, vol. 2, no. 5, pp. 425–436, 2013.

[25] S. K. Zahid, L. Hasan, A. A. Khan, and S. Ullah, "A novel structure of the smith-waterman algorithm for efficient sequence alignment," in *2015 Third International Conference on Digital Information, Networking, and Wireless Communications (DINWC)*, pp. 6–9, IEEE, 2015.

[26] A. S. Konagurthu, J. C. Whisstock, P. J. Stuckey, and A. M. Lesk, "Mustang: a multiple structural alignment algorithm," *Proteins: Structure, Function, and Bioinformatics*, vol. 64, no. 3, pp. 559–574, 2006.

[27] J. Shoji, H. Hinoo, T. Kato, T. Hattori, K. Hirooka, K. Tawara, O. Shiratori, and Y. Terui, "Isolation of cepafungins i, ii and iii from pseudomonas species," *The Journal of antibiotics*, vol. 43, no. 7, pp. 783–787, 1990.

[28] M. Beckwith, W. J. Urba, and D. L. Longo, "Growth inhibition of human lymphoma cell lines by the marine products, dolastatins 10 and 15," *JNCI: Journal of the National Cancer Institute*, vol. 85, no. 6, pp. 483–488, 1993.

[29] C. A. Müller, J. Markovic-Lipkovski, T. Klatt, J. Gamper, G. Schwarz, H. Beck, M. Deeg, H. Kalbacher, S. Widmann, J. T. Wessels, *et al.*, "Human α-defensins hnps-1,-2, and-3 in renal cell carcinoma: Influences on tumor cell proliferation," *The American journal of pathology*, vol. 160, no. 4, pp. 1311–1324, 2002.

[30] M. H. Abo-Ghalia, G. O. Moustafa, A. E.-G. E. Amr, A. M. Naglah, E. A. Elsayed, and A. H. Bakheit, "Anticancer activities of newly synthesized chiral macrocyclic heptapeptide candidates," *Molecules*, vol. 25, no. 5, p. 1253, 2020.

[31] A. E.-G. E. Amr, R. E. A. Mageid, M. El-Naggar, A. M. Naglah, E. S. Nossier, and E. A. Elsayed, "Chiral pyridine-3, 5-bis-(l-phenylalaninyl-l-leucinyl) schiff base peptides as potential anticancer agents: Design, synthesis, and molecular docking studies targeting lactate dehydrogenase-a," *Molecules*, vol. 25, no. 5, p. 1096, 2020.

[32] G. O. Moustafa, A. Shalaby, A. M. Naglah, M. M. Mounier, H. El-Sayed, M. M. Anwar, and E. S. Nossier, "Synthesis, characterization, in vitro anticancer potentiality, and antimicrobial activities of novel peptide–glycyrrhetinic-acid-based derivatives," *Molecules*, vol. 26, no. 15, p. 4573, 2021.

[33] O. F. Lamidi, M. Sani, P. Lazzari, M. Zanda, and I. N. Fleming, "The tubulysin analogue kemtub10 induces apoptosis in breast cancer cells via p53, bim and bcl-2," *Journal of cancer research and clinical oncology*, vol. 141, pp. 1575–1583, 2015.

[34] S. Ran, A. Downes, and P. E. Thorpe, "Increased exposure of anionic phospholipids on the surface of tumor blood vessels," *Cancer research*, vol. 62, no. 21, pp. 6132–6140, 2002.

[35] Y. Liscano, J. Oñate-Garzón, and J. P. Delgado, "Peptides with dual antimicrobial–anticancer activity: Strategies to overcome peptide limitations and rational design of anticancer peptides," *Molecules*, vol. 25, no. 18, p. 4245, 2020.

[36] S. Vijayakumar and L. Ptv, "Acpp: a web server for prediction and design of anti-cancer peptides," *International Journal of Peptide Research and Therapeutics*, vol. 21, pp. 99–106, 2015.

[37] W. Chen, H. Ding, P. Feng, H. Lin, and K.-C. Chou, "iacp: a sequence-based tool for identifying anticancer peptides," *Oncotarget*, vol. 7, no. 13, p. 16895, 2016.

[38] S. Akbar, M. Hayat, M. Iqbal, and M. A. Jan, "iACP-GAEnsC: Evolutionary genetic algorithm based ensemble classification of anticancer peptides by utilizing hybrid feature space," *Artificial Intelligence in Medicine*, vol. 79, pp. 62–70, 2017.

[39] L. Xu, G. Liang, L. Wang, and C. Liao, "A novel hybrid sequence-based model for identifying anticancer peptides," *Genes*, vol. 9, no. 3, 2018.

[40] C. Wu, R. Gao, Y. Zhang, and Y. De Marinis, "PTPD: predicting therapeutic peptides by deep learning and word2vec," *BMC Bioinformatics*, vol. 20, p. 456, Sept. 2019.

[41] A. Gupta, M. Chauhan, R. Tomer, and G. Raghava, "Anticp3: Prediction of anticancer proteins using evolutionary information from protein language models," 2025. preprint, not peer-reviewed.

[42] M. Pirtskhalava, A. A. Amstrong, M. Grigolava, M. Chubinidze, E. Alimbarashvili, B. Vishnepolsky, A. Gabrielian, A. Rosenthal, D. E. Hurt, and M. Tartakovsky, "DBAASP v3: database of antimicrobial/cytotoxic activity and structure of peptides as a resource for development of new therapeutics," *Nucleic Acids Research*, vol. 49, pp. D288–D297, Jan. 2021.

[43] S. Singh, K. Chaudhary, S. K. Dhanda, S. Bhalla, S. S. Usmani, A. Gautam, A. Tuknait, P. Agrawal, D. Mathur, and G. P. Raghava, "Satpdb: a database of structurally annotated therapeutic peptides," *Nucleic Acids Research*, vol. 44, pp. D1119–D1126, 11 2015.

[44] A. Shendre, N. K. Mehta, A. S. Rathore, N. Kumar, S. Patiyal, and G. P. S. Raghava, "Map format for representing chemical modifications, annotations, and mutations in protein sequences: An extension of the fasta format," 2025.

[45] L. Yang, S. Cao, L. Liu, R. Zhu, and D. Wu, "cyclicpeptide: a python package for cyclic peptide drug design," *Briefings in Bioinformatics*, vol. 26, p. bbae714, 01 2025.

[46] S. A. González-Chávez, S. Arévalo-Gallegos, and Q. Rascón-Cruz, "Lactoferrin: structure, function and applications," *International Journal of Antimicrobial Agents*, vol. 33, no. 4, pp. 301.e1–301.e8, 2009.

[47] F. Lu, Y. Zhu, G. Zhang, and Z. Liu, "Renovation as innovation: Repurposing human antibacterial peptide LL-37 for cancer therapy," *Frontiers in Pharmacology*, vol. 13, Aug. 2022. Publisher: Frontiers.

[48] P. Agrawal and G. P. S. Raghava, "Prediction of Antimicrobial Potential of a Chemically Modified Peptide From Its Tertiary Structure," *Frontiers in Microbiology*, vol. 9, Oct. 2018. Publisher: Frontiers.

[49] V. Kumar, R. Kumar, P. Agrawal, S. Patiyal, and G. P. S. Raghava, "A Method for Predicting Hemolytic Potency of Chemically Modified Peptides From Its Structure," *Frontiers in Pharmacology*, vol. 11, Feb. 2020. Publisher: Frontiers.